RISK IDENTIFICATION AND ASSESSMENT OF HUMAN-MACHINE CONFLICT

by

© He Wen

A thesis submitted to the School of Graduate Studies

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Faculty of Engineering and Applied Science

Memorial University of Newfoundland

May 2023

St. John's Newfoundland and Labrador

Abstract

The process industries are fully embracing digitalization and artificial intelligence (AI). Industry 4.0 has also transformed the production structures in the process industries to increase productivity and profitability; however, this has also led to emerging risks. The rapid growth and transformation have created gaps and challenges in various aspects, for example, information technology (IT) vs. operation technology (OT), human vs. AI, and traditional statistical analysis vs. machine learning.

A notable issue is the apparent differences in decision-making between humans and machines, primarily when they work together. Contradictory observations, states, goals, and actions may lead to conflict between these two decision-makers. Such conflicts have triggered numerous catastrophes in recent years. Moreover, conflicts may become even more elusive and confusing under external forces, e.g., cyberattacks.

Therefore, this thesis focuses on human-machine conflict. Five research tasks are conducted to explore the risk of human-machine conflict. More specifically, the thesis presents a systematic literature review on the impact of digitalization on process safety, highlights the myths and misconceptions of data modeling on process safety analysis, and attempts to clarify associated concepts in the area of human-machine conflict. In addition, the thesis summarizes the causes of conflicts and generalizes the mathematical expressions of the causes. It illustrates the evolutional process of conflicts, proposes the measurement of conflicts, develops the risk assessment model of conflicts, and explores the condition of conflict convergence, divergence, and resolution. The thesis also

demonstrates the proposed methodology and risk models in process systems, for example, the two-phase separator and the Continuous Stirred Tank Reactor (CSTR). It verifies the conflict between manual and automated control (e.g., proportional-integral-derivative control (PID) and model predictive control (MPC)).

This thesis proves that conflict is another more profound and implicit phenomenon that raises risks more rapidly and severely. Conflicts are highly associated with faults and failures. Various factors can trigger human-machine conflict, including sensor faults, cyberattacks, human errors, and sabotage. This thesis attempts to provide the readers with a clear picture of the human-machine conflict, alerts the industry and academia about the risk of human-machine conflict, and emphasizes human-centered design.

Acknowledgment

First, I would like to thank my supervisor Dr. Faisal Khan sincerely. I started and finished my doctoral research during the COVID-19 pandemic. It is Dr. Khan who gave me this opportunity to return to academia, helped me realize my childhood dream, and changed my life in the past three years and my career in the future.

Dr. Khan's profound knowledge and rigorous academic philosophy have deeply nurtured and influenced me. I still remember those frank discussions, trustful communication, tireless explanations, rational debates, and complementary cooperation, often even at midnight on weekends and holidays. I am truly blessed to be supervised by such a scholar as Dr. Khan.

During this academic journey, I also genuinely thank my co-supervisors, Dr. Syed Imtiaz and Dr. Salim Ahmed. I received the foremost help from them in any situation. Their creative research thoughts, critical comments, motivational feedback, and infinite patience support me and this thesis. In addition, I wholeheartedly thank my senior, Dr. Md. Tanjin Amin, who guided my study and collaborated on my research. And I sincerely appreciate Dr. Stratos Pistikopoulos and Dr. Syeda Z. Halim for collaborating on my research.

I also honestly thank Dr. Bing Chen, Dr. Yahui Zhang, Colleen Dalton, Tina Dwyer, Nicole Parisi, Jinghua Nie, Andrew Kim, Jennifer Kennedy, and Ruby Barron, who have provided research support and coordinated my academic activities.

At last, I would like to thank my family, the ones I love. My parents care about my

health, safety, and living thousands of miles away, and I hope this thesis can comfort their ardent hopes. I also thank my partner Fengxiang, who accompanied me to study at night when we were thousands of miles apart and took care of my diet and daily life when we got together, so that I could do my research with peace of mind. Thank you for your love and dedication.

Table of Contents

Lists of Figures

Lists of Tables

Lists of Abbreviations

ACC: adaptive cruise control

ADAS: advanced driver assistance system

AHP: analytic hierarchy process

AI: artificial intelligence

ANN: artificial neural network

BN: Bayesian network

BT: bow tie

CACE: Computers & Chemical Engineering

CART: classification and regression tree

CDF: cumulative density function

CPT: conditional probability table

CSTR: continuous stirred tank reactor

DoS: denial of service

DT: decision tree

ETA: event tree analysis

EU-OSHA: European Agency for Safety and Health at Work

FDI: false data injection

FT: fuzzy theory

FTA: fault tree analysis

HMI: human machine interface; human machine interaction

ICA: independent component analysis

ICS: industrial control system

IEC: International Electrotechnical Commission

IoT: Internet of Things

ISO: International Organization for Standardization

IT: information technology

JHM: Journal of Hazardous Materials

JLPPI: Journal of Loss Prevention in the Process Industries

JRR: Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability

JSR: Journal of Safety Research

KMC: k-means clustering

KNN: k-nearest neighbor

LCL: lower control limit

LOPA: layer of protection analysis

LQR: linear quadratic regulator

LR: logistic regression

MCAS: maneuvering characteristics augmentation system

ML: machine learning

MPC: model predictive control

MSPM: multivariate statistical process monitoring

M-ANN: myth of ANN (overuse of ANN)

M-BN: myth of BN (missing BN's underlying assumption)

M-CC: myth of correlation coefficient (using correlation coefficient for model verification)

M-DR: myth of data representation (improper data representation)

M-EA: myth of error analysis (absence of error analysis)

M-MSPM: myth of multivariate statistical process monitoring (absence of model behavior analysis for MSPM)

NBC: naïve bayes classifier

OT: operation technology

PCA: principal component analysis

PDF: probability density function

PID: proportional-integral-derivative controller

PLC: programmable logic controller

PLS: partial least square

PN: Petri net

PSE: process system engineering

PSEP: Process Safety and Environmental Protection

PSP: Process Safety Progress

RA: Risk Analysis

RESS: Reliability Engineering & System Safety

RF: random forest

SA: situation awareness

SS: Safety Science

SVM: support vector machine

UCL: upper control limit

VAD: variable of action difference

VID: variable of interpretation difference

VOD: variable of observation difference

WoS: Web of Science Core Collection

Lists of Symbols

a: a constant related to the flow rate out of the tank; the half range of the adjustable input; subscript of interpretation

A: state matrix

b: a constant related to the flow rate into the tank; subscript of interpretation

B: input matrix

C: output matrix; cross-sectional area of the tank

C_A: concentration of the reagent in the reactor

C_{Af}: concentration of the reagent in the inlet feed stream

d: difference of observation or action; distance between the vector of AI interpretation probability and the vector of human interpretation probability

D: feedthrough matrix; Laplace transform of the difference function

d_u: variable of action difference (VAD)

d_x: variable of observation difference (VOD)

e: a random error

$e(t)$: a changing error

f: function from observation to action

f_0: function from true value to observation

f_{A0}: function from true value to sensor observation

f_{H0}: function from true value to human observation

f_1: function from observation to interpretation

f_{A1}: function from observation to interpretation of AI

f_{H1}: function from observation to interpretation of human

f_2: function from interpretation to action

F: cumulative density function

g: function from observation to action

G: transfer function of the controller

h: height of oil in the tank

H: transfer function of the process

J: loss function

k: order or the Pareto point

K_u: coefficient constant of the valve opening

m: observation vector size

M: achievable matrix in loss function

n: action vector size; the transformed interpretation vector size

N: normal distribution

O: achievable matrix in the loss function

P: conflict probability

q: inlet flow rate

Q: achievable matrix in the loss function

r: setpoint

$\tilde{r}$: modified setpoint

R: reference function; conflict risk

R_{max}: the maximum risk at the upper and lower limit of input

s: complex variable

S: conflict severity

t: time

T: temperature in the reactor

T_c: temperature of the jacket coolant

T_f: temperature of the inlet feed stream

u: action or input; valve opening

$\bar{u}$: input at the setpoint

$\hat{u}$: human action or human input; most possible human action

$\tilde{u}$: input in hacked status; most possible AI action

u_C: controller action

u_H: human control action

U: input variable or action in the s domain

$\widetilde{U}$: hacked input variable or action in the s domain

v: generalized attack vector

V: volume of oil in the tank; generalized attack in the s domain

w: attack vector or attack method

W: attack in the s domain

x: state variable or sensor observation

$\tilde{x}$: observation in hacked status; most possible sensor observation

$\hat{x}$: human observation or expectation; most possible human observation

x_C: sensor observation

$\hat{x}_C$: sensor observation without a fault

x_H: human observation

y: output variable; observation as the output variable; interpretation

$\hat{y}$: most possible human interpretation

$\tilde{y}$: most possible AI interpretation

Y: output function

α: parameter of a Beta distribution

β: parameter of a Beta distribution

$BETA.INV$: return the inverse of the beta cumulative probability density function

σ: standard deviation.

Δ: a constant

ε: a very small positive number

μ: mean

η: noise

η_1: noise in sensor observation

η_2: noise in AI interpretation

η_3: noise in human observation

η_4: noise in human interpretation

ω: true value

τ: delayed time

Chapter 1: Introduction

1.1. Background and motivation

Industry 4.0, or digitalization, is an umbrella concept to describe a fusion of technologies. Common digital technologies include artificial intelligence (AI), machine learning (ML), big data, simulation, system integration, autonomous robotics, additive manufacturing, the Internet of Things (IoT), cloud computing, cybersecurity, and augmented reality (Figure 1.1) (Rüßmann et al., 2015). Undoubtedly, digitalization integrating information technology (IT) and operation technology (OT), has radically transformed the production structures, procedures, and operations in the process industries (Pistikopoulos et al., 2021), with higher productivity and profitability (Arunthavanathan et al., 2020; Nian et al., 2020). However, it also raises challenges, conflicts, myths, and misconceptions and widens gaps in understanding process operations (Bécue et al., 2021; Khan et al., 2021). Such emerging risks have triggered notable accidents.

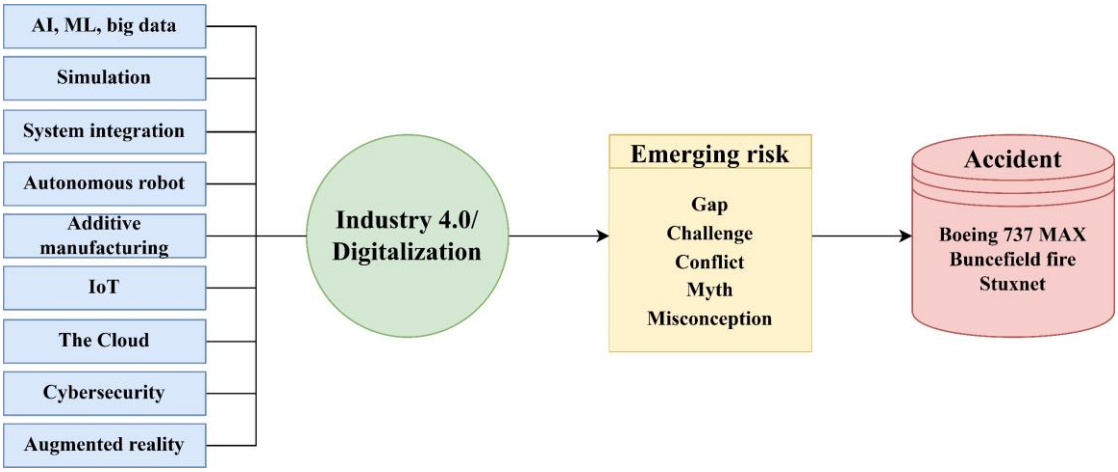

Figure 1.1: Digitalization and emerging risks.

Such an often-cited example is the Boeing 737 MAX accident (DeFazio & Larsen, 2020). Though the causes of the crash are complicated, one of the root causes is the incorrect data given by the single angle of attack sensor. Consequently, the autopilot forced the plane into a dive and conflicted with the pilot's correction.

The second example is the Buncefield fire (Buncefield Major Incident Investigation Board, 2008); the cause of this accident is mainly attributed to a gauge failure. However, the human operator witnessed similar observation conflicts as the pilot in the case of Boeing 737 Max crash.

The third example is Stuxnet which attacked the programmable logic controllers (PLC) and caused damage to centrifuges in a nuclear plant (Chen, 2010; Kushner, 2013). The malware forced the centrifuge to overspin, while the operator did not recognize the reason. This is a confusing phenomenon, and any operator action may be in conflict with the tampered control action, thus, resulting in a conflict between the operator and the control system.

The above accidents demonstrate the roles humans and machines play in highly automated and intelligent systems, which may lead to new challenges to process operations, such as conflicts between humans and machines. Therefore, this thesis aims to explore the gaps and challenges in process digitalization, more specifically, the human-machine conflict.

1.2. Objectives and tasks

The main objective of this thesis is to explore emerging risks of digitalization with a

specific focus on the human-machine conflict. Five tasks are performed sequentially (Figure 1.2) to meet the objectives.

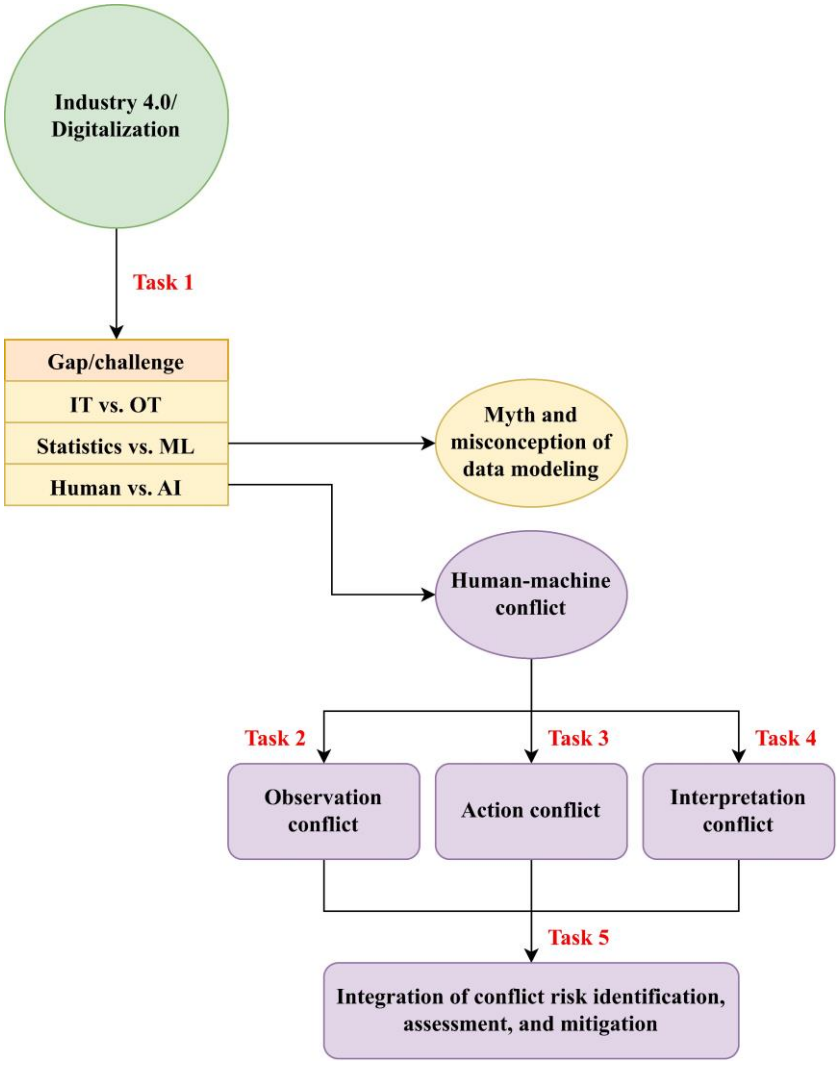

Figure 1.2: Research flowchart.

In the beginning, a systematic literature review (task 1) is conducted, with the keywords "process safety" and "digitalization, automation, artificial intelligence, machine learning, data mining, Industry 4.0, smart manufacturing". Thus, three significant gaps and challenges are identified – IT vs. OT, human vs. AI, and traditional statistical analysis vs. machine learning. Consequently, human-machine conflict is selected as the

research focus and explored in three directions – observation conflict (task 2), action conflict (task 3), and interpretation conflict (task 4). Task 5 is an integrated study based on cyberattacks to illustrate and apply the proposed concepts and risk models. The detailed objectives and corresponding tasks are summarized in Table 1.1.

Table 1.1: Research objectives and tasks.

Objective	**Task**
To identify gaps and challenges related to digitalization's impact on safety	Task 1: Conduct a systematic literature review
To explore the nature of conflict and observation conflict during process system digitalization	Task 2: Define conflict variables to explore conflict evolution and resolution
To assess and manage the action conflict during security threat	Task 3: Develop a game paradigm to express conflict due to cyberattack
To analyze the interpretation conflict between humans and AI during critical operation	Task 4: Propose a distance variable to measure interpretation conflict
To unify the approach to manage emerging conflict risks during process system digitalization	Task 5: Perform an integration study

This thesis attempts to provide the reader with a clear picture of the human-machine conflict, alerts the industry and academia about the risk of human-machine conflict, and emphasizes human-centered design.

1.3. Outcomes and novelties

The significant contributions of this thesis are the novel concept of human-machine conflict and the mathematical illustrations. The research outcomes and novelties of this thesis are presented in Table 1.2.

Table 1.2: Outcomes and novelties.

Task	Outcomes	Novelties and key contributions
Task 1	Myths and misconceptions of data-driven methods: Applications to process safety analysis. *Computers and Chemical Engineering*, (2022), 158, 107639	● Summarized gaps and challenges of digitalization. ● Presented six most frequent myths and misconceptions of data modeling. ● Conducted a systematic review of human-machine conflict and associated topics.
Task 2	A methodology to assess human-automated system conflict from safety perspective. *Computers & Chemical Engineering*, (2022), *165*, 107939	● Proposed the concept of human-machine conflict and mathematical properties. ● Demonstrated the evolutional nature of a conflict. ● Developed a novel methodology to assess the conflict risk.
Task 3	Risk Assessment of Human-Automation Conflict under Cyberattacks in Process Systems. *Computers & Chemical Engineering*, (2023), 172, 108175	● Extended the human-machine conflict under attack. ● Proposed attack representation with process variables and parameters. ● Presented a mathematical formulation of conflicts with the game paradigm.
Task 4	Assessment of Situation Awareness Conflict Risk between Human and AI in Process System Operation. Submitted to *Industrial & Engineering Chemistry Research*, (2023), 62(9)	● Proposed the distance between humans and AI to measure conflict. ● Explored the impacts of various noises on conflicts.
Task 5	-	● Performed an integration study of conflict risk identification, assessment, and mitigation

1.4. Outline of thesis

The outline of this thesis and research framework is presented in Figure 1.3.

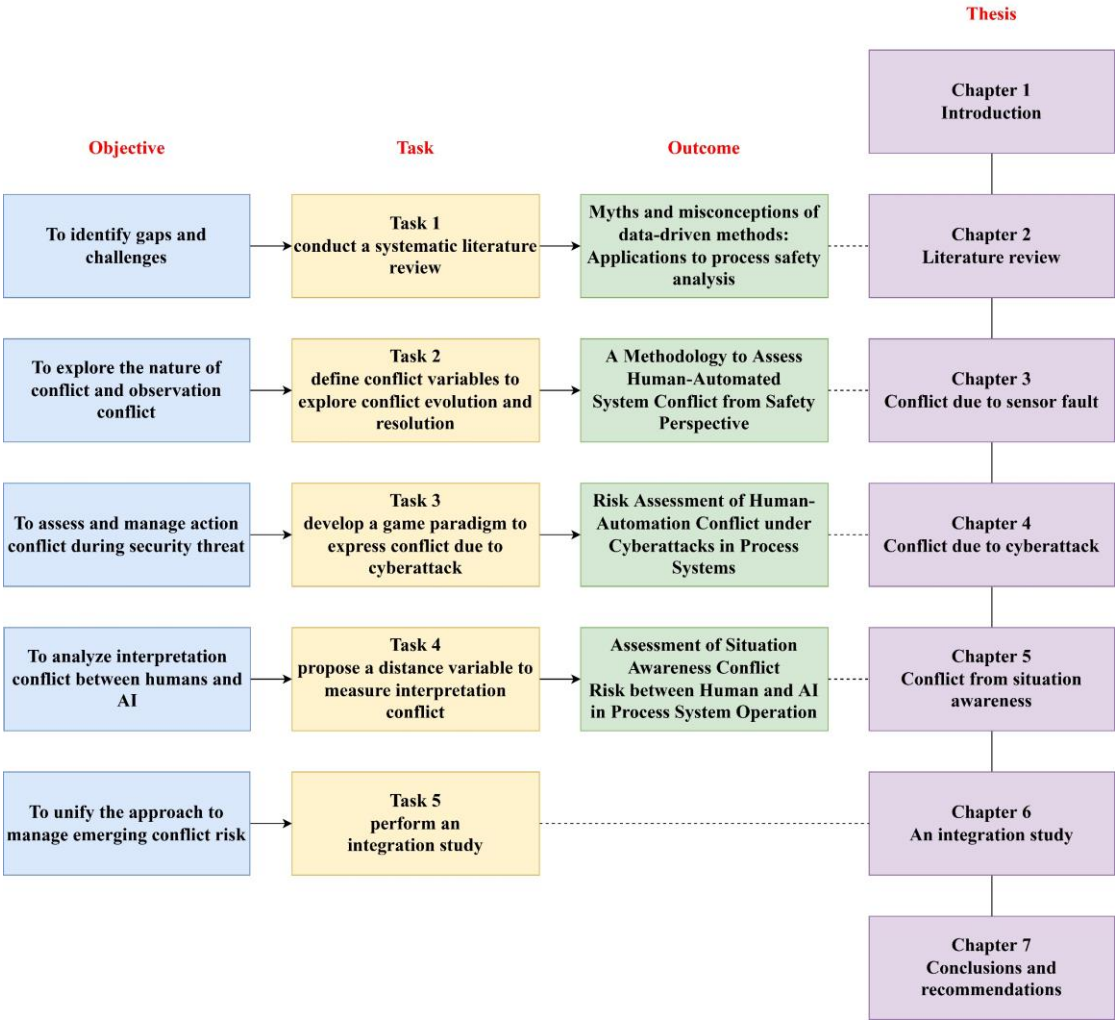

Figure 1.3: Thesis outline and research framework.

1.5. Co-authorship statement

I am the primary author of this thesis under the direct supervision of Dr. Faisal Khan and the co-supervision of Dr. Salim Ahmed and Dr. Syed Imtiaz. This thesis comprises five main research tasks presented in Chapters 2-6. These chapters offer research papers published in peer-reviewed journals. I am the primary author of these research papers. To achieve the best outcomes for the research tasks, I have collaborated with Dr. Stratos Pistikopoulos, Dr. Md. Tanjin Amin, and Dr. Syeda Z. Halim. I conducted the literature

review and developed the methodology and its applications. I have prepared the first draft of the manuscript and subsequently revised it based on the co-authors' comments. Dr. Faisal Khan formulated the project, granted the research question, and helped me develop the research concept, methodology, and models. Drs. Salim Ahmed, Syed Imtiaz, Stratos Pistikopoulos, Md. Tanjin Amin and Syeda Z. Halim contributed to preparing, reviewing, and revising the manuscript.

Chapter 2: Literature Review

Preface

This chapter provides a literature review relevant to the thesis; part of this chapter is published in *Computers & Chemical Engineering*. I am the primary author. I have conducted this work under the direct supervision of Dr. Faisal Khan, who also co-authored this work. I have collaborated on this with Md. Tanjin Amin, and Syeda Z. Halim. I developed the conceptual framework for the paper and carried out the literature review. I prepared the first draft of the manuscript and subsequently revised the manuscript based on the co-authors' and peer review feedback. Co-author Dr. Faisal Khan assisted in developing the conceptual model, research methodology, analysis of results, reviewing, and revising of the manuscript. Co-authors Drs. Md. Tanjin Amin and Syeda Z. Halim supported implementing the concept. The co-authors provided fundamental assistance in validating, reviewing, and correcting the model and results. The co-authors also contributed to the review and revision of the manuscript.

Reference: Wen, H., Khan, F., Amin, M. T., & Halim, S. Z. (2022). Myths and misconceptions of data-driven methods: Applications to process safety analysis. *Computers and Chemical Engineering*, 158, 107639. https://doi.org/10.1016/j.compchemeng.2021.107639

2.1. Gaps and challenges of digitalization

Admittedly, digitalization may raise new risks. Professional organizations, such as International Electrotechnical Commission (IEC) (The International Electrotechnical

Commission [IEC], 2020) and European Agency for Safety and Health at Work (EU-OSHA) (European Agency for Safety & Health at Work [EU-OSHA], 2018), have summarized and forecasted the impact of digitalization on safety. Shared viewpoints include zero goal, human-centered safety, collaborative safety, and remote work issues. A bibliometric search was conducted from the Web of Science Core Collection (WoS) database to obtain a holistic view of digitalization's impact on process safety. The searching keywords were “process safety” and “digitalization, automation, artificial intelligence, machine learning, data mining, Industry 4.0, smart manufacturing”. The timespan for this search was set between 1990 and 2022. Articles were downloaded from Memorial University’s library and analyzed in detail.

Significant gaps and challenges are identified; three are notable – IT vs. OT, human vs. AI, and statistical analysis vs. machine learning. First is the fragmentation and isolation of IT/OT (Ehie & Chilton, 2020; Garimella, 2018), or in other words, the convergence difficulty of the cyber system and physical system (Kamal et al., 2016; Paes et al., 2020). This involves the gap between operators' capability and digital technology requirements (Khan et al., 2021). Moreover, system convergence introduces cyber threats (Iaiani, Tugnoli, Bonvicini, et al., 2021a).

The Second is the gap between humans and AI, or the level difference between idealized intelligence and realistic automation (Peres et al., 2020; Vagia et al., 2016). Achievable intelligence is still far away from full automation. Nevertheless, this expands the discussion between humans and the automated system, for example, the impact on

human performance (Endsley & Kaber, 1999; Kaber & Endsley, 2004; Park et al., 2019), situation awareness of large-scale systems (Naderpour et al., 2015), process operation problems with intelligent systems (Benson et al., 2021), authority and priority issues (Inagaki, 2003; Tessier & Dehais, 2012), assignment of roles and tasks (Frohm et al., 2008), cooperation and competition (Briken, 2020; Parasuraman et al., 2000).

The third is the leap between traditional statistical analysis and AI/ML-based data mining (Bzdok et al., 2018; Mirkin, 2011). Traditional knowledge discovery mainly relies on statistics and logic deduction; however, current machine learning emphasizes correlation. This makes researchers pay less attention to internal logic and causal relationship. In addition, scholars from different disciplines may have a limited specific understanding of algorithms and models, which raises myths and misconceptions about data modeling.

Therefore, two critical issues are worthy of attention in this literature review. One is the myth and misconception of data modeling, and the other is human-machine conflict.

2.2. Myth and misconception of data modeling

2.2.1. Analysis methodology

This work consists of six major steps, as shown in Figure 2.1. The steps are discussed below.

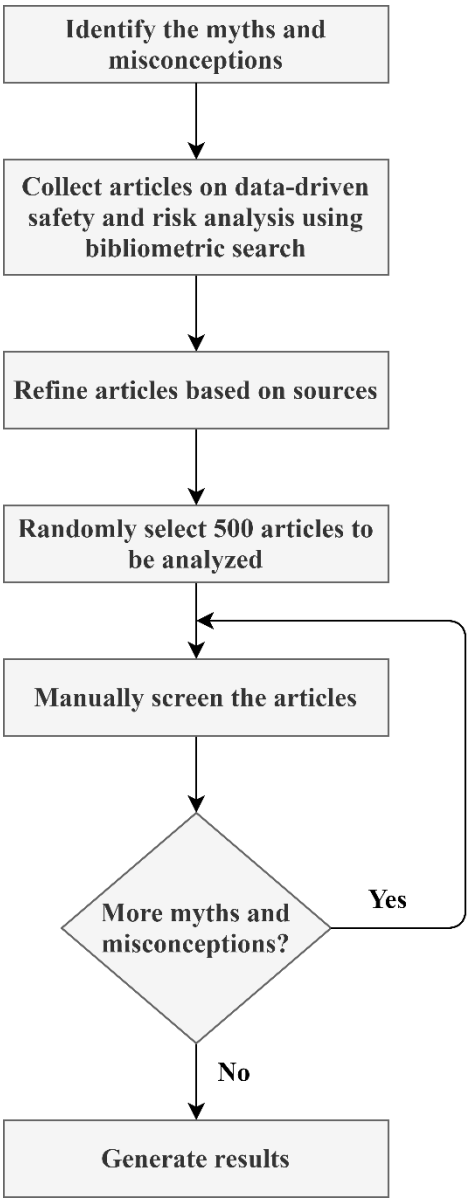

Figure 2.1: Flowchart of the analysis methodology.

Step 1: At the beginning of this study, the authors discussed the noticed myths and misconceptions in data-driven methods. As a result of this brainstorming, improper data representation, Absence of model behavior analysis for multivariate statistical process monitoring (MSPM), missing Bayesian network's (BN) underlying assumption, and overuse of artificial neural network (ANN) were shortlisted as the frequently observed myths. The subsequent steps were conducted to measure the frequency of these myths.

Step 2: In this step, a total of 19 frequently used data-driven methods, such as the fault tree analysis (FTA), event tree analysis (ETA), bow tie (BT), Bayesian network, fuzzy theory (FT), analytic hierarchy process (AHP), Petri net (PN), artificial neural network, support vector machine (SVM), decision tree (DT), random forest (RF), classification and regression tree (CART), naïve Bayes classifier (NBC), k-means clustering (KMC), k-nearest neighbor (KNN), logistic regression (LR), principal component analysis (PCA), independent component analysis (ICA), and partial least squares (PLS), were selected to collect documents.

A bibliometric search was conducted from the WoS database on May 28th, 2021. The timespan for this search was set from 1990 to 2020. This search resulted in 808,266 articles. Although there are some other renowned databases, such as Scopus and Compendex, this work only collected bibliographic data from the WoS database. Unlike the other two databases, WoS stores lesser conference proceedings. However, it contains articles from a wide range of fields. Since the scope of this work is to scrutinize the journal articles, WoS was used. The searching technique was as follows:

TOPIC: *fault tree analysis* OR *event tree analysis* OR *bow tie* OR *Bayesian network* OR *fuzzy theory* OR *analytic hierarchy process* OR *Petri net* OR *artificial neural network* OR *support vector machine* OR *decision tree* OR *random forest* OR *classification and regression tree* OR *naïve Bayes* OR *k-means* OR *k nearest neighbor* OR *logistic regression* OR *principal component analysis* OR *independent component analysis* OR *partial least squares*

Step 3: The bibliometric search generated a large number of articles since they were collected from a wide variety of fields and sources. To narrow down the number of articles for in-depth analysis and to include the articles most relevant to process safety, articles from ten renowned safety journals (Table 2.1) were screened out.

These journals were Reliability Engineering & System Safety (RESS), Computers & Chemical Engineering (CACE), Safety Science (SS), Journal of Hazardous Materials (JHM), Journal of Loss Prevention in the Process Industries (JLPPI), Risk Analysis (RA), Process Safety and Environmental Protection (PSEP), Journal of Safety Research (JSR), Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability (JRR), and Process Safety Progress (PSP). This resulted in 3,290 articles.

Table 2.1: Source-wise collected samples.

Journal	Original sample	Randomly selected sample	Percent selected (%)
RESS	835	150	30.0%
CACE	563	102	20.4%
SS	437	53	10.6%
JHM	322	27	5.4%
JLPPI	314	53	10.6%
RA	242	31	6.2%
PSEP	217	50	10.0%
JRR	142	17	3.4%
JSR	140	4	0.8%
PSP	78	13	2.6%
Total	3290	500	100.0%

Step 4: The articles obtained from the previous step were further narrowed down to 500 documents using a random search technique. Since the WoS gave the result in a tab-delimited format, the output was not readily available for a random selection. The tab-delimited file was converted to an excel file. All the articles were sorted out in

alphabetical order. A random number, ranging from 1 to 3,290, was then assigned to each article as its identifier. Finally, random numbers were generated 500 times to select the final manuscripts to be analyzed for this research. The goal of this random search is to reduce the bias in article selection.

Table 2.1 shows the number of articles collected from each source, while Table 2.2 displays the number of articles found from each data-driven method. It should be noted that the total number of randomly selected papers is seen as 748. The same article can contain the use of two or more methods. This is the reason for a larger sample size than the actual one.

Table 2.2: Method-wise collected samples.

Method	Original sample	Randomly selected sample	Percent selected (%)
FTA	517	115	15.4%
ETA	419	90	12.0%
BT	115	27	3.6%
BN	599	127	17.0%
FT	647	108	14.4%
AHP	144	18	2.4%
PN	177	31	4.1%
ANN	327	40	5.3%
SVM	140	33	4.4%
DT	290	50	6.7%
RF	50	7	0.9%
CART	29	4	0.5%
NBC	16	5	0.7%
KMC	36	4	0.5%
KNN	13	4	0.5%
LR	251	10	1.3%
PCA	386	54	7.2%
ICA	101	9	1.2%
PLS	135	12	1.6%
Total	4392	748	100.0%

Step 5: After getting the list of 500 manuscripts, we downloaded them from Memorial

University's library. At this stage, we looked for the frequency of the type of data source and myths and misconceptions. We manually read these articles in this context. After analyzing 500 articles, a total of five data sources were identified: measurement, historical database, survey, expert knowledge, and simulation. In the analysis, two other myths were observed to occur commonly: using correlation coefficient for model verification, and absence of error analysis. Subsequently, the analysis was reiterated to determine the frequency of occurrence of these newly identified myths and misconceptions. It was also noticed that most ANN-based documents do not provide any rationale behind selecting the hyperparameters that are crucial for ANN's performance. These myths were defined as follows and included in the search list (Table 2.3).

The definitions of the identified myths and misconceptions are:

1) Improper data representation (M-DR): Violation of standard rules on significant digit, arithmetic calculation, or data uncertainty, which are stipulated in standards by the International Organization for Standardization (ISO) or other recognized institutions. This study analyzed four subcategories: digit inconsistency, inaccurate calculation to significant digits, false precision, and improper uncertainty.
2) Absence of model behavior analysis for MSPM (M-MSPM): Ignorance of the prerequisites when applying MSPM tools. This study analyzed two subcategories: no examination of linear or nonlinear data pattern, and no identification of variable distribution.

3) Missing BN's underlying assumption (M-BN): Lack of reasonable assumptions on parameter dependence within models and conditional probability tables (CPTs). This study analyzed two subcategories: misuse for overall dependence modeling, and unclear CPTs.

4) Overuse of ANN (M-ANN): Unreasonable alternative research for simple scenarios which could be interpreted by mature physical and chemical laws, or setting self-defined parameters and arbitrary parameter adjustments. This study analyzed two subcategories: using ANN to replace simple analytical equations and setting arbitrary hyper parameters.

5) Using correlation coefficient for model verification (M-CC): Misunderstanding the correlation coefficient, and misusing it for model verification, which is due to garble with the coefficient of determination. This study analyzed one myth: using correlation coefficient for model verification.

6) Absence of error analysis (M-EA): Lack of error analysis indicators to evaluate the model performance, or no clear error justification or discussion for the model and the research. This study analyzed one myth: absence of error analysis.

Table 2.3: Commonly noticed myths and misconceptions.

Myth/misconception	Abbreviation	Subcategory
Improper data representation	M-DR1	Digit inconsistency
	M-DR2	Inaccurate calculation to significant digits
	M-DR3	False precision
	M-DR4	Improper uncertainty
Absence of model behavior analysis for MSPM	M-MSPM1	No examination of linear or nonlinear data pattern
	M-MSPM2	No identification of variable distribution
Missing BN's underlying assumption	M-BN1	Misuse for overall dependence modeling
	M-BN2	Unclear CPTs
Overuse of ANN	M-ANN1	Using ANN to replace simple analytical equations
	M-ANN2	Setting arbitrary hyper parameters
Using correlation coefficient for model verification	M-CC	Using correlation coefficient for model verification
Absence of error analysis	M-EA	Absence of error analysis

Step 6: Finally, the frequency of articles for each type of data source, myth, and misconception was counted, and their statistical representation was provided.

2.2.2. Key results

The 500 articles were analyzed to numerically list the number of samples that contain at least one myth. The result summary is displayed in Table 2.4. The total number of articles containing such myths and misconceptions is 168, which is 33.6% of the analyzed samples, and 288 cases were found. The biggest share is caused by being less attentive to data representation (163 cases). Another crucial observation is the use of BN without stating proper assumptions (55 cases).

Table 2.4: Statistical summary of myths and misconceptions.

Myth/misconception	Subcategory	Frequency	Frequency
M-DR	M-DR1	121	163
	M-DR2	13	
	M-DR3	8	
	M-DR4	21	
M-MSPM	M-MSPM1	9	19
	M-MSPM2	10	
M-BN	M-BN1	31	55
	M-BN2	24	
M-ANN	M-ANN1	12	32
	M-ANN2	20	
M-CC	-	10	10
M-EA	-	9	9
Total		288	288

The method-wise frequency of considered myths and misconceptions was also analyzed (Figure 2.2). The BN-based articles contain the largest portion (96 cases), followed by the MSPM-based (49 cases) and ANN-based (43 cases) articles. As can be seen from Figure 2.2, the proportion of these myths and misconceptions in BN-based articles was significantly higher than in other methods and far outweighed the proportion of randomly chosen BN-based articles to others in Table 2.2. Except for the BN-based articles, M-DR is the dominant myth in other method-based documents. Absence of proper assumption is mostly noticed in BN-based articles.

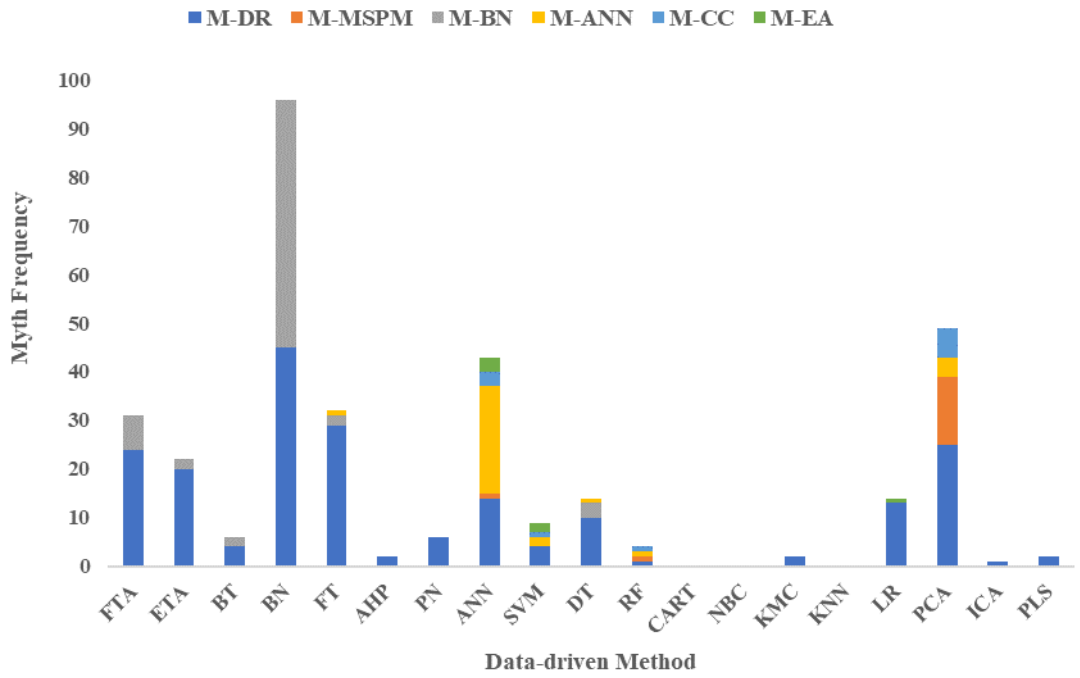

Figure 2.2: Method-wise myths and misconceptions.

2.3. Conflict and human-machine conflict

2.3.1. Conflict

The definition of "conflict" from a linguistic source is *an antagonistic state or action* (*Conflict | Definition of Conflict by Merriam-Webster*, n.d.). It involves multiple participants with different value systems in a state of disagreement. Nevertheless, the conflict has distinct explanations in Social Science, Psychology, and Computer Science. In the social psychological study of conflict, multiple illustrations of "conflict" have been proposed. One is widely accepted that *conflict is a process in which one party perceives that its interests are being opposed or negatively affected by another party* (Wall & Callister, 1995). This involves fundamental discussions on interpersonal communication, cooperation, competition, individual perception, attitude, and task

orientation (Deutsch, 1990).

A widely accepted view in Computer Engineering, the conflict has three conditions: *i) the agents have at least two contradictory goals, ii) the agents are aware of their goals to be contradictory, and iii) the agents have to make a choice* (Castelfranchi, 2000). In the human-machine system, primarily based on the pilot-autopilot system, a further developed definition is that *a conflict is the execution of actions that are effective but in spite of this are either logically incoherent, either physically incoherent or epistemically incoherent* (Pizziol, 2013). Thus, the basis of a conflict refers to its linguistic meaning, with the common understanding that a conflict is a contradictory state or action between multiple participants.

Among multiple levels and types of conflict (S. M. Easterbrook et al., 1993), this thesis focuses on one of them, which is the human-human conflict. It won't involve the discussion of the mechanism of psychology, neither the political level of ethnic conflict nor conflict of war. Fortunately, typical causes of conflict have been concluded (Wall & Callister, 1995), which can be benchmarked with human-machine conflict (Table 2.5). This has particular reference significance.

Table 2.5: Causes of conflict and the difference.

Cause of human-human conflict	Difference in human-machine conflict
Personality	Machines do not have personalities.
Values	Machines do not have value judgment.
Goals	Goals may be different.
Stress, anger	Machines have no psychological factors.
Communications and interaction	Communication is one-way from human to machine, not vice versa.
Distrust	Humans may not trust machines.
Misunderstanding	Humans may misunderstand the situation.
Reduction of other's outcomes	Machines eliminate human interference
Power imbalances	Who has priority and authority?
Vague vs. Clear	Humans can handle ambiguous tasks, while machines cannot.

Furthermore, it develops strategies to win the conflict and how to resolve or de-escalate it. The noted and practical one is the "two-dimensional taxonomy of conflict-handling modes" (Thomas, 1992), and five strategies are presented: avoiding, competition, compromise, accommodation, and collaboration. This will inspire Human-machine conflict resolution.

2.3.2. Human-machine conflict

In-depth studies of conflict are rare in process industries, as the focus is more on fault diagnosis and abnormal situation management. The aviation industry has been applying autopilot systems for decades. Hence, it has accumulated experience in such situations of human-machine conflict. The common phenomena are known as "automation surprise" (Dehais et al., 2015) or "mode confusion" (Hamburger, 1966; Leveson et al., 1997). Similar situations are hesitation, doubt, and unsureness.

Moreover, conflicts have already evolved into catastrophes. The well-known Boeing 737 Max crashes are notable examples in this regard (DeFazio & Larsen, 2020). One of

the fundamental reasons is that the maneuvering characteristics augmentation system (MCAS) pushed the aircraft into a dive due to incorrect data from the angle-of-attack sensor. Further, the automated system prevented the pilot from recovering the aircraft manually. This is a typical sensor fault, triggering observation conflict, then interpretation conflict, and action conflict. However, sensor fault is often emphasized here rather than human-machine conflict. Similarly, problems of situation awareness conflict are often blamed on the inappropriate design of automation (Sarter & Woods, 1991). Yet, it becomes even more profound when automation extends to AI because humans have a strong sense of a situation, while AI does not.

Traditionally, human-machine conflict is studied from the perspective of the human factor, which assumes that the autopilot is entirely correct while human behavior deviates (Dehais et al., 2003) or exhibits anomalies of situation awareness (Endsley, 1995; Endsley & Kaber, 1999). Further, human responses to mechanical system failure are studied (Beringer & Harris, 1999; Woods & Sarter, 1998). Undoubtedly, human error can lead to conflict in this state (Dehais et al., 2015). Yet, what needs to be reflected here is whether it is the system failure that causes the conflict or whether the conflict is caused by human error. The answer may be the combination. Therefore, one solution is detecting, predicting, and modeling automated surprise and mode confusion (Bredereke & Lankenau, 2005; Hamburger, 1966). The other solution is to improve the reliability of automated systems (Leveson et al., 1997) and discuss the priority and authority of humans and machines to avoid such conflicts (Inagaki, 2003).

In the automotive industry, self-driving cars provide abundant situations of conflicts. Phantom braking, unexpected acceleration, or lane change confuses the driver (Moscoso Paredes et al., 2021), and then the driver may have situational reactions under stress. According to a survey, over 70% of the respondents experienced phantom braking at least once under different driving speeds and conditions (Moscoso Paredes et al., 2021). Authority management and conflict resolution are also the frontier topic for self-driving cars (Tessier & Dehais, 2012).

In addition, cyberattacks may also be supposed to trigger such conflicts. For example, Stuxnet is the first malware targeting PLC (Chen, 2010; Kushner, 2013; Langner, 2013). The malware turned the centrifuges overspin to tear apart. Though the operator may be aware of the abnormality, it is yet arduous to resolve it. Another example is that a hacker accessed the human-machine interface (HMI) of a water plant in the USA and tampered with the level of sodium hydroxide from 100 ppm to 11,100 ppm (Campo-Flores, 2021). A peculiar phenomenon is that the value was changed repeatedly when the operator corrected it.

Conflict resolution involves deciding on final authority and priority, or the leader and follower in multi-agents. (Inagaki, 2003) demonstrated an automated system could provide a safer consciousness than a human and discussed how to prioritize human and automated system actions. However, researchers have mentioned that human-centered design should be stressed (Boy, 2017; Shneiderman, 2021). Furthermore, to improve the performance and resolve the conflict is also studied from the perspective of HMI

design (J. Li & Vachtsevanos, 2014), collaborative system design (X. Li et al., 2002), and computer-supported negotiation (S. Easterbrook, 1991; S. M. Easterbrook, 1994).

2.3.3. Human-machine relationship

When expanding the human-human conflict to the human-machine conflict, it is necessary to discuss the relationship between humans and machines first. A review paper constructs metrics to measure human operators and machines (Damacharla et al., 2018), presenting the associated topics, such as human-machine teaming, human-machine interaction, human-machine interface, and human-machine cooperation. After the benchmark, Figure 2.3 and Table 2.6 summarize correlated issues and their association.

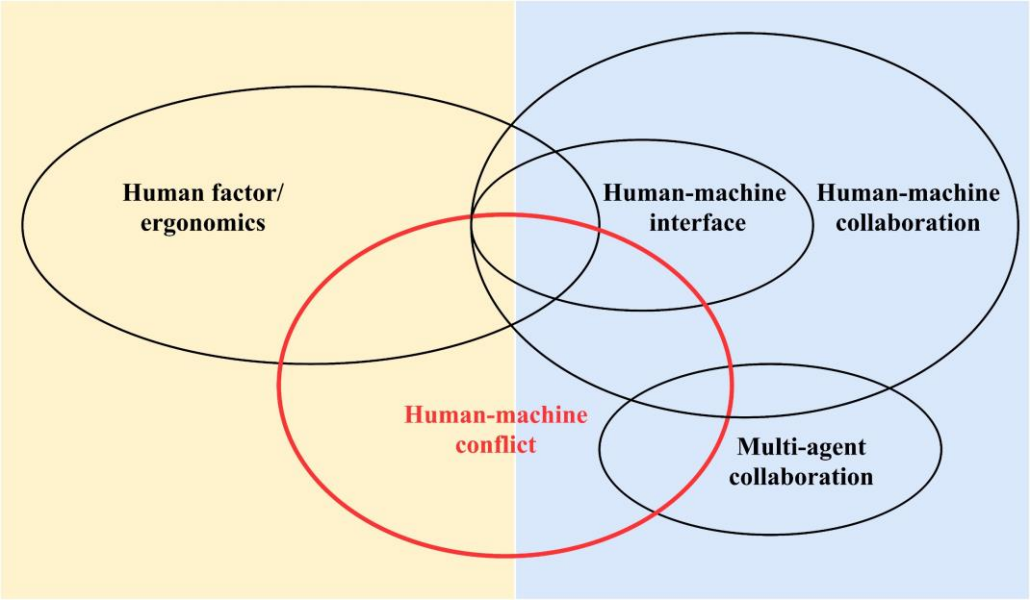

Figure 2.3: Human-machine relationship and related topics.

Table 2.6: Related topics in the human-machine relationship.

Type	Who	What	How
Human-human conflict	Human and other humans	Act for different goals	Independent actions
Human-machine conflict	Human and machine	Different observations, interpretations, and actions on the same task	Humans monitor machines and assist them
Human factor and ergonomics	Human and machine	Ensure productivity and safety	Independent work aiming to reduce human error
Human-machine Interaction	Human and machine/computer, or their interface	Humans and machines communicate and interact through user interfaces	Do their part in the common task
Human-machine collaboration	Human and machine/AI	Work synchronously for the shared task	Do their part in the common task and even complete each other's unfinished parts
Multi-agent conflict	Smart machine/AI	Work synchronously for the shared task	Do their part in the shared task
Human-AI ethical conflict	Humans and AI	Ethics and Survival	Fight for survival

- **Human factor and ergonomics**

Human factor research started in France in the 1930s (Hollnagel, 2018), and it has developed numerous practical assessment methods and human error databases. According to the definition of International Ergonomics Association (International Ergonomics Association, n.d.), *ergonomics (or human factors) is the scientific discipline concerned with the understanding of interactions among humans and other elements of a system, and the profession that applies theory, principles, data and*

methods to design to optimize human well-being and overall system performance.

Both ergonomics and human-machine conflict explore the relationship between humans and machines. They also involve the study of cognition and behavior. When humans and machines cooperate, humans may make mistakes. Ergonomics discusses how humans use machines and tools correctly and efficiently to avoid human error, which aims to maintain productivity. In the meantime, it helps safety performance in disguise. Humans are still the weaker link in an ever-improving system's reliability, and the operator's role is critical (Edwards & Lees, 1971).

Another problem is that human error research is usually separated from automated system failure, as researchers try to focus on specific scenarios for human error. Analysis methods, for instance, layer of protection analysis (LOPA), consider human intervention to deal with the alarm without extension on human error analysis (Baybutt, 2002; Dowell, 1998; Myers, 2013).

Human errors are usually unintentional, while violation and sabotage are not. Violation is intentionally breaking the rules and acting on the machines against the operating procedures. Moreover, sabotage is when humans deliberately destroy machines. Admittedly, human error, violation, and sabotage may also cause conflicts.

- **Human-machine interaction**

As an extension of ergonomics, human-machine interaction (HMI) studies from cognitive, behavioral, and human capacity perspectives (Vinciarelli et al., 2015; Wang et al., 2021), also in conjunction with control engineering (Vogel et al., 2015). It is

mainly in the interaction between interfaces, called the human-machine interface. Later, it gradually refers to the human-computer interaction or the human-computer interface. Similar studies also extend to human-robot interaction.

Research in this area includes the performance evaluation of HMI or user image/user experience (Ha and Seong, 2009), human reliability in HMI (Strand and Lundteigen, 2017), and HMI failure (Sudano and Marietta, 1994). Besides Computer Science, self-driving and autopilot are critical scenarios of human-machine interaction, integrating multiple tasks and environmental changes (Shan, 2021).

- **Human-machine collaboration**

Furthermore, research extends from human-machine interaction to human-machine cooperation (Hoc, 2000), as HMI serves human-machine collaboration. Since humans and machines work together frequently, especially when machines become more automated and intelligent, the cooperation relationship has been reformed. Specifically, the relationship is classified as co-existence, cooperation, and collaboration (Aaltonen et al., 2018; Kolbeinsson et al., 2019; Magrini et al., 2020; Vicentini, 2020). The discussions are more about the collaborative relationship (Flemisch et al., 2019) and even mutual trust collaboration (Alhaji et al., 2020; Visser et al., 2018).

Nowadays, it refers more to collaborative robotics, especially robotic arms. The research includes, for example, exploring how the division of labor between humans and robotic arms in the same co-located cell more efficiently and safely (Ferreira et al., 2021; Magrini et al., 2020). Moreover, functional safety research on robotic arms

(Fryman & Matthias, 2012; Salmi et al., 2012; Zanchettin et al., 2016) also extends to new concepts of collaborative safety (Mukaidono et al., 2018). Admittedly, in the autonomous driving scenario, the relationship between human-machine collaboration is also discussed in more depth (C. Huang et al., 2021; Vanderhaegen, 2021; Weyer et al., 2015).

In addition to the joint one-on-one human-machine collaboration, humans and machines cooperate more closely in the case of human-machine teaming. In machine-dominated systems, the human-in-the-loop problem is often considered (Cohen & Singer, 2021; Inoue et al., 2019; H. N. Wu et al., 2021). In a human-dominated system, machines will be considered agents joining the team. Further research includes team form and efficiency in human-machine teaming (Stowers et al., 2021; Walliser et al., 2019), primarily on how to design AI systems to make teams more efficient (Saenz et al., 2020) and the work intensity of humans (Heard & Adams, 2019).

Nevertheless, as AI spreads in industry applications, the interaction between humans and AI becomes familiar, and collaboration safety problems increase dramatically. In contrast, the combination of AI failure and human error has not drawn enough attention. One combination is the human in the loop control system, which is to design the human as a component in the closed-loop control system (Cohen & Singer, 2021) or to study the human intervention and its influence (H. N. Wu et al., 2021). To quantify the risk between humans and AI, the first thing to do is to study the collaboration relation. In industrial applications, autonomous vehicles (Aptiv et al., 2019), collaborative robots

(Aaltonen et al., 2018; Magrini et al., 2020), and medical devices (Freschi et al., 2013) have exhibited leading-edge exploration. Unfortunately, the types of human-AI collaboration in process safety remain unidentified, and the risk of human-AI collaboration needs to be quantified.

- **Multi-agent conflict**

The multi-agent system can still complete specific tasks without human intervention. In the case of multi-agent teaming, the problem of multi-agent computing conflict is highlighted (Canonico, 2019). Computational conflicts are instruction conflicts caused by multiple agents and multiple computing nodes (Castelfranchi, 2000). At a macro level, this topic includes multi-agent conflict and cooperation in the field of control engineering, in particular, multi-agent control may lead to collision problems (Y. Huang et al., 2020; Zhou, 2021) and optimal path planning problems (Sharon et al., 2015; Zero et al., 2019).

- **Human-AI ethical conflict**

In specific tasks, the computational power of deep learning algorithms far exceeds that of humans, such as in the game of Go (Silver et al., 2016, 2017). This kind of AI applied in games is playing a competitive match against human intelligence across the board (Westera et al., 2020). Human-computer gaming is the comparison and competition of the intellect and computing power of the two. In addition, the adversarial machine learning spawned by adversarial attacks helps AI consider human judgment in the learning process, making AI more intelligent, accurate, and reliable.

This has raised concerns about whether AI will entirely surpass or even replace humans. The ethical conflict between humans and AI involves these issues (Coombs et al., 2021; Malle et al., 2019), for example, whether AI will have self-awareness, whether it will resist humans, and whether it will destroy humans. At least for now, AI is taking human jobs. There are also issues such as the ethical conflict of driverless accidents, how to set ethical choices, and the conflict of moral choices between humans and autonomous vehicles. Furthermore, it involves the research of AI security, for example, the malicious use of AI, the deceptiveness of generative AI, and how to achieve reliable and friendly AI. Significantly, human-centered AI should be emphasized in design (Shneiderman, 2021).

2.4. Identified knowledge gaps

From the above literature review, knowledge gaps are identified:

- The issue of human-machine conflict has received noticeable attention; however, it has not been considered in combination with humans and machines. In the past, it was viewed separately from the perspective of ergonomics or system failure.
- There is neither complete and widely accepted definition of human-machine conflict nor detailed introduction and classification of the phenomena and situations of human-machine conflict.
- It also lacks the classification of conflicts, such as observation conflict, interpretation conflict, and action conflict, as well as the difference between them. There is also a lack of mathematical expression and derivation for them.

- The causes of the conflict have not been deeply researched and understood, and hence, further generalizations and summaries are needed.
- In the process of deep integration of IT and OT, process engineers often lack adequate IT knowledge. They do not have a good understanding of the hazards brought by new threats, such as cyberattacks. The resulting conflict phenomenon is more difficult to understand and requires in-depth research.
- Practitioners are using countless AI applications and constantly improving the way and accuracy of AI imitating humans. Still, the differences and gaps between humans and AI are the root causes of the risks of current AI applications. The gaps and risks should be measured and assessed.
- For conflict resolution, there have been remarkable conceptual studies. However, they are not yet practical, mainly since the mathematical conditions for conflict resolution have not been well defined.

Chapter 3: Conflict Due to Sensor Fault

Preface

A version of this chapter has been published in *Computers & Chemical Engineering*. I am the primary author, along with the co-authors, Drs. Faisal Khan, Md. Tanjin Amin, Salim Ahmed, Syed Imtiaz, and Stratos Pistikopoulos. I developed the conceptual framework for the methodology and carried out the literature review and case study. I prepared the first draft of the manuscript and subsequently revised the manuscript based on the co-authors' and peer review feedback. Co-author Dr. Faisal Khan helped develop the concept, verify the methodology, review, and revise the manuscript. Co-authors Drs. Md. Tanjin Amin, Salim Ahmed, Syed Imtiaz, and Stratos Pistikopoulos provided support in implementing the concept and verifying the methodology. The co-authors provided fundamental assistance in validating, reviewing, and correcting the methodology, case study, and results. The co-authors also contributed to the review and revision of the manuscript.

Reference: Wen, H., Amin, M. T., Khan, F., Ahmed, S., Imtiaz, S., & Pistikopoulos, S. (2022). A methodology to assess human-automated system conflict from safety perspective. *Computers & Chemical Engineering*, 165, 107939. https://doi.org/10.1016/j.compchemeng.2022.107939

Abstract

Automated systems have exhibited enormous prospects in applications. Most automated systems are equipped with shared control systems with two intelligent decision-makers:

humans and automated machines. The contradictory observations, states, goals, and actions may result in a conflict between these two decision-makers. The definitions, cause(s), and path(s) of such a conflict from a process safety perspective have not been explored and assessed in the existing literature. This work introduces an evolutionary framework that shows how a conflict can lead to an accident. A methodology and associated models to assess and manage conflict risk are also presented. The methodology and models are explained using a two-phase separator. The results suggest that there are conflicts (i.e., observations and actions) associated with faults that may lead to failure. A sensor fault can trigger observation conflict, which may lead to action conflict. The study concludes that human-automated system conflict in automation and digitalization should be emphasized. Human-centered design is vital to avoid catastrophic accidents due to conflicts in human-automated systems.

Keywords: Conflict, fault diagnosis, failure analysis, probabilistic risk analysis, human-automated system, digitalization.

3.1. Introduction

In the past few decades, the applications of Artificial Intelligence (AI) in process industries have seen geometric growth (J. Lee et al., 2019; Muhuri et al., 2019). It is paving the way for digital technologies to be employed in automated control systems, or short for automated systems. As a result, process plants are embracing digitalization and digital transformation at a rapid pace (Klatt & Marquardt, 2009; Pistikopoulos et al., 2021). This trend is benefitting the process industries with increased profit and fewer

failures (Arunthavanathan et al., 2020; Nian et al., 2020; Pistikopoulos et al., 2021). The most remarkable benefit has been experienced during the ongoing COVID-19 pandemic when digitalization has enabled process plants to be operated without the physical presence of operators (Acioli et al., 2021; Oliva et al., 2021). Regrettably, increased dependence on digital technologies has brought new challenges and threats (Gobbo et al., 2018; Khan et al., 2021). One of the challenges is the rising conflict risk when humans and automated machines work synchronously (Briken, 2020). For an intelligent industrial product, one of the usual forms of AI is the automated system. Before AI completely replaces human operators, there exist two intelligent decision-makers: the human operator and the automated machine. This combination is the source of conflict.

In our daily life, self-driving cars can be considered to illustrate the conflict. Automated vehicles provide easier driving, parking, and timesaving. However, the upgraded automation level may result in conflict between the driver and the car. A common phenomenon is “phantom braking” (Moscoso Paredes et al., 2021), which is when the vehicle brakes unexpectedly due to the interference of the advanced driver assistance system (ADAS). According to a survey, over 70% of the respondents experienced phantom braking at least once in their lives, under different driving speeds and different conditions (Moscoso Paredes et al., 2021).

In the aviation industry, this phenomenon is known as “automation surprise” (Dehais et al., 2015) or “mode confusion” (Hamburger, 1966; Leveson et al., 1997). Though the

aviation industry has been applying autopilot systems for decades, numerous catastrophes have occurred due to such conflict scenarios. The well-known Boeing 737 Max crashes are notable examples in this regard (DeFazio & Larsen, 2020). One of the fundamental reasons is that the maneuvering characteristics augmentation system (MCAS) pushed the aircraft into a dive due to faulty data from the angle-of-attack sensor. Further, the automated system prevented the pilot from recovering the aircraft manually.

The level of automation in process industries is significantly lower than in the aviation and automotive sectors; therefore, the conflict scenarios are fewer than in these two industries. However, due to the severity of a catastrophic accident in process industries, in-depth research on conflict analysis is required to avoid catastrophic failures caused by a conflict.

The definition of "conflict" from a lexical source is *an antagonistic state or action* (*Conflict | Definition of Conflict by Merriam-Webster*, n.d.). It involves multiple participants with different value systems, and they are still in a state of disagreement. Nevertheless, the conflict has distinct explanations in Social Science, Psychology, and Computer Science. A widely accepted view in Computer Engineering, the conflict has three conditions: *i) the agents have at least two contradictory goals, ii) the agents are aware of their goals to be contradictory, and iii) the agents have to make a choice* (Castelfranchi, 2000). In the human-automated system, especially based on the pilot-autopilot system, a further developed definition is that *a conflict is the execution of*

actions that are effective but in spite of this are either logically incoherent, either physically incoherent or epistemically incoherent (Pizziol, 2013). Thus, the basis of a conflict refers to its linguistic meaning, with the common understanding that a conflict is a contradictory state or action between multiple participants.

In process systems, a conflict may arise due to disagreement between the human operator and the automated machine, specifically, the control system driven by AI algorithms or automation. Therefore, the conflict or human-automated system conflict in the process industries is the difference in the observation, interpretation, or action of one or more variables by different participants (the human operator and the automated system). In this sense, a conflict is a condition of disagreement between two sets of information or action. Observation conflict would be more likely to arise, for example, the sand in crude oil may contaminate and trigger the malfunction of the level sensor in oil-gas separation, and the sensor reading is often different from the operator's observation. This work focuses on action conflict that may be driven by conflicting observation or interpretation of the observation.

In-depth studies of conflict are rare in process industries. However, a few documents are available in the context of the aviation and auto industries. (Damacharla et al., 2018) reviewed the metrics to measure human operators and machines, presenting the associated topics, such as human-machine teaming, human-machine interaction, human-machine interface, and human-machine cooperation. Nevertheless, human-automated system conflict is relatively different from the above topics, as the conflict

may result in unharmonized scenarios and risks. Multiple studies have shown that conflict is highly related to the cognition difference between human operators and automated systems. Interviews with experienced drivers on their behaviors in cut-in scenarios showed the significant difference in cognitive and behavioral patterns between the drivers and the adaptive cruise control (ACC) led to conflicts (Gong et al., 2019). Simulations of specific conflict scenarios have shown how the drivers responded to the conflict (Pipkorn et al., 2021). Both humans and automated systems have positive attributes. Automated systems are more strictly compliant with legal requirements, while humans have better philosophical judgment and the ability to adapt based on real-time information.

In addition, game-theoretic approaches for autonomous vehicles have presented remarkable research between the driver and ADAS in cooperative and conflict scenarios, for example, the decision-making in steering control (X. Li & Wang, 2021), velocity control (K. Huang et al., 2020), lane change (Sankar & Han, 2020; Zhang et al., 2019), and game paradigms between the driver and the control system (Na & Cole, 2015, 2017). More discussions have surpassed the limitation of two participants, such as the interactions among multiple drivers and vehicles (N. Li et al., 2018), the multi-agent system control (Canonico, 2019; Jost et al., 2017), and the team awareness and conflict (McNeese et al., 2021).

Conflict resolution involves the discussion on final authority and priority, the leader and follower in multi-agents. (Inagaki, 2003) showed an automated system could provide a

safer consciousness than a human and discussed how to prioritize human and automated system actions. However, researchers have mentioned that human-centered design should be stressed (Boy, 2017; Shneiderman, 2021). The prerequisite of conflict resolution is to quantify the conflict and generate the conventional transforming procedure. Hence, mathematical simulation should be applied and strived. (Pizziol et al., 2014) applied the Petri net to model and simulate the conflict between the pilot and the aircraft; however, the conflict scenario considered in this work is narrowed and exclusive and cannot be expanded to generic applications.

The above discussion suggests that research on conflict has not gained attention in process industries. Therefore, it is worthwhile to assess the risk of a conflict in the context of process plants, and it will be an essential step of industrial safety analysis with the increasing adoption of Industry 4.0 and smart controllers. The current work is undertaken with this motivation. It presents a detailed analysis of the evolution of a conflict. Also, a framework is presented for quantitative conflict risk assessment and conflict resolution. The contributions of this work are:

(i) Introduction of a novel concept of human-automated system conflict in process safety assessment;

(ii) Demonstration of evolutional nature of a conflict;

(iii) New definitions and mathematical properties of conflict variables;

(iv) Development of mathematical expressions for conflict risk assessment.

The reminder of this article is organized as follows: Section 3.2 describes the distinct

steps of the proposed methodology for conflict risk assessment and management; an application of this framework to a two-phase separator is discussed in Section 3.3; the advantages, limitations, and future work scopes are discussed in Section 3.4; the concluding remarks are summarized in Section 3.5.

3.2. Methodology to identify and assess conflicts

3.2.1. Research flowchart

This study consists of two parts: methodology to access human-automated system conflict and demonstration of its application to a two-phase separator. The different steps involved in the study are shown in Figure 3.1, while a brief description is provided below.

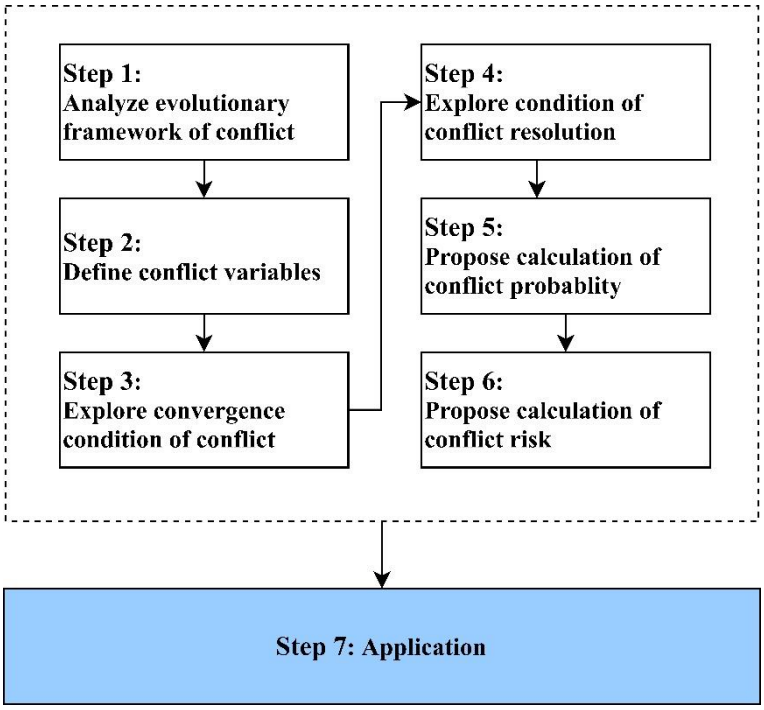

Figure 3.1: Details of the steps involved in the present study.

Step 1: At the beginning of this study, the relation between fault, failure, and conflict

was discussed, and an evolutionary framework of observation conflict and risk conflict was proposed.

Step 2: To quantify the conflict, two variables were introduced and defined clearly.

Step 3: The convergence and divergence conditions of conflict were expressed mathematically.

Step 4: The conditions of conflict resolution were identified according to the convergence expression. Furthermore, the human intervention was summarized to present how the conflict can be resolved.

Step 5: Conflict probability was discussed and proposed with mathematical equations.

Step 6: The calculations of conflict severity and conflict risk were proposed.

Step 7: A two-phase separator was described and introduced to apply the above steps. A simulation by MATLAB/Simulink R2021a was conducted and the conflict risk was calculated.

3.2.2. Conflict evolution

In process safety analysis, a conflict is based on a fault, as decision-makers (i.e., human and automated system) will usually be involved when a fault occurs. Consider a sensor measurement, and it is accurately measured by the instruments. Therefore, any significant deviation or fault will properly be captured. However, it may be interpreted differently by the human operator and automated control system. Suppose a deviation in a process variable generates an alarm. The controller will take action accordingly; nonetheless, the operator may respond differently since flooded alarms are considered

a nuisance in process plants. Therefore, there will be an action difference that may result in a conflict. If this conflict is not resolved, it may cause a failure.

Phenomenally, a conflict is declared a failure, while a fault is the symptom of the conflict. Fundamentally, both conflict and failure are resultants of a fault (Figure 3.2). Nevertheless, it can be seen that there is no failure when some conflicts occur, for example, observation conflict arises phantom braking without any failure. Moreover, action conflict may trigger a mechanical failure of the actuator. Therefore, conflict is one of the sources of failure. This work focuses on the overlapping area between a conflict and a failure. A special focus is given to model how a conflict will lead to failures and accidents.

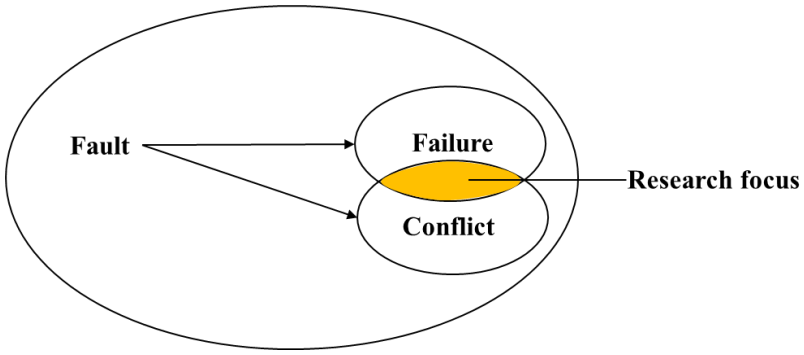

Figure 3.2: The relation between fault, failure, and conflict.

Based on the relation between fault, failure, and conflict, the evolutionary framework for the conflict risk assessment and management is proposed in Figure 3.3. The procedures and scenarios are discussed below.

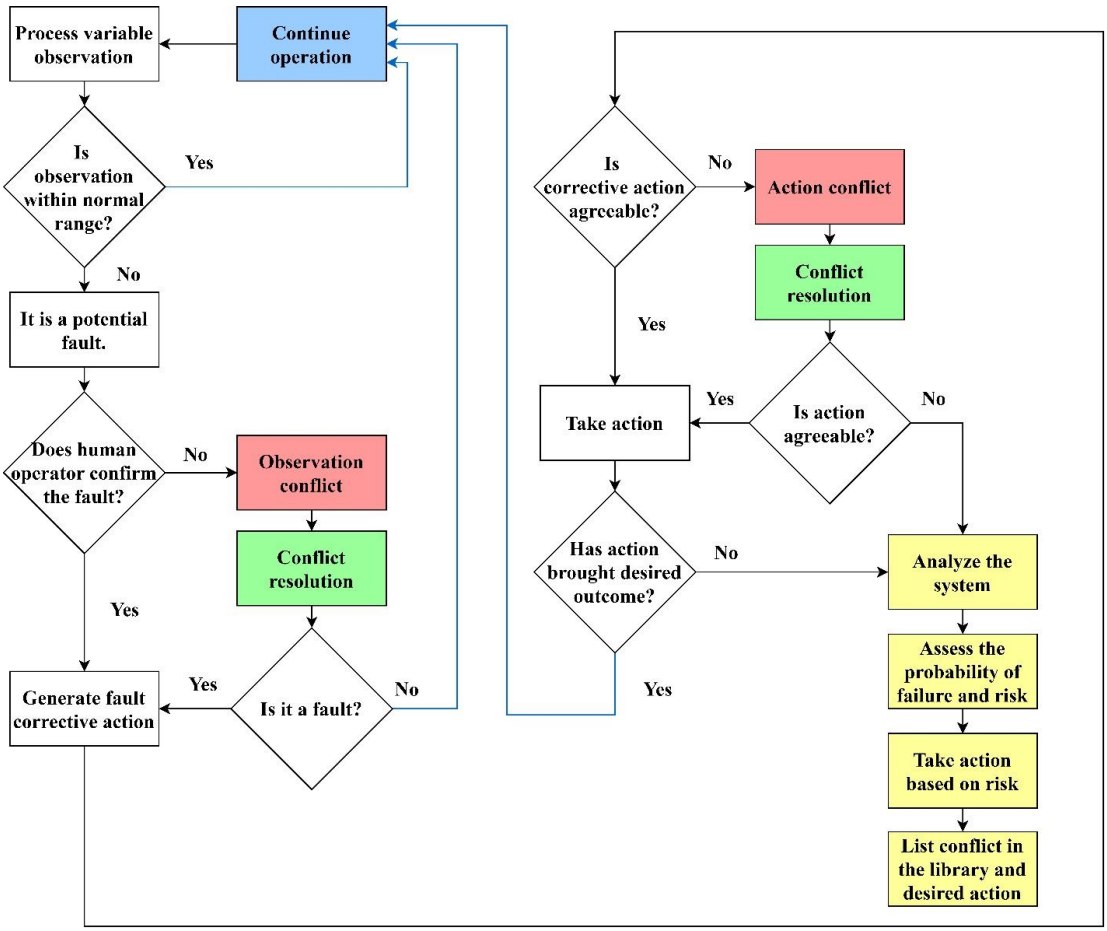

Figure 3.3: Evolutionary framework for conflict risk assessment.

Scenario 1: In most cases without faults, the sensor detects the process variable, and the sensor observation is maintained in the normal range, then the system continues operation.

Scenario 2: Once the sensor detects a fault, it will generate an alarm. The human operator is alerted to confirm the fault. If the human observation cannot confirm the sensor observation, an observation conflict occurs. The human operator will intervene to resolve the conflict. After the human intervention, if it is no longer a fault, the system continues operation.

Scenario 3: If the system is still at fault after human intervention on the observation

conflict, further fault corrective action will be generated. When the corrective actions are agreeable, the automated system or the human operator will take action. Once the desired outcome is reached, the system continues operation.

Scenario 4: When the human operator and the automated system do not agree on the corrective action, an action conflict occurs. After the human intervention, the action conflict can be resolved. If the automated system or the human operator agrees on the corrective action, they will take action. Once the desired outcome is reached, the system continues operation.

Scenario 5: Or the desired outcome cannot be reached, the system goes into risky operation and needs risk assessment and management.

Scenario 6: The last scenario is that the action conflict cannot be resolved and the corrective action on the fault is still not agreeable. The system goes into risky operation and needs risk assessment and management.

3.2.3. Conflict variables

3.2.3.1. Variable of observation difference

The variable of observation difference (VOD) is the difference in observation of process value from different observers. Suppose sensor observation of automated control system as $x_C(t)$, human observation as $x_H(t)$, and VOD as $d_x(t)$.

$$d_x(t) = x_C(t) - x_H(t) \tag{3.1}$$

Usually, process data follow Gaussian distribution and suppose $N(\mu, \sigma^2)$, where N stands for normal distribution, μ is the mean, and σ is the standard deviation. At first,

an observation conflict indicates high chances of a fault (Figure 3.4), which means $x_C \notin [\mu - 3\sigma, \mu + 3\sigma]$ or $x_H \notin [\mu - 3\sigma, \mu + 3\sigma]$.

The expectation of human observation is $E(x_H) = x_C = \mu$, then VOD follows $d_x \sim N(0, \sigma^2)$. A reasonable range of human observation is $x_H \in [x_C - \sigma, x_C + \sigma]$, which means a 68.2% possibility that the observation difference is not significant (Montgomery & Runger, 2010).

Consequently, the judgment condition of observation conflict is $d_x \notin [x_C - (x_C + \sigma), x_C - (x_C - \sigma)]$, which is $d_x \notin [-\sigma, \sigma]$. The lower control limit (LCL) and upper control limit (UCL) of VOD are $-\sigma$ and σ, respectively (Figure 3.4).

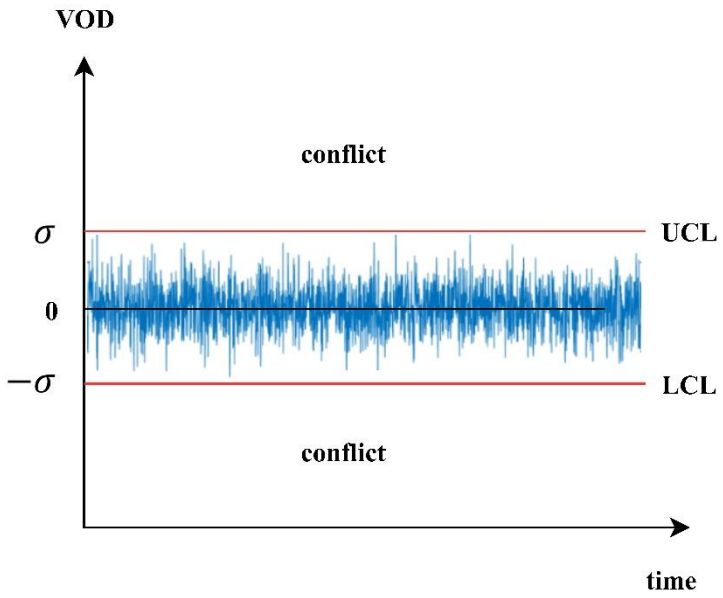

Figure 3.4: VOD and observation conflict.

Observation conflict is mostly due to sensor fault or human error. If the human operator relies on another sensor, both fault types are sensor faults. Redundant sensors would reduce the failure rate of the sensing system; consequently, the possibility of observation

conflict would be reduced indirectly. On the other side, as human error is difficult to predict and quantify accurately, this paper omits discussion on it. Common sensor faults are shown in Table 3.1 (Yung & Clarke, 1989). VODs are shown in Figure 3.5 based on Equation (3.1), supposing the faults occur at time 100 s in the schematic diagrams. This paper focuses on drift fault which represents the variability of different faults.

Table 3.1: Sensor fault types and mathematical expressions.

Sensor fault type	**Mathematical expression**
Short-circuit	$x_C(t) = 0$
Open-circuit	$x_C(t) = \infty$
Stuck	$x_C(t) = \hat{x}_C(t_0)$, $\hat{x}_C$ is sensor observation without a fault.
Bias	$x_C(t) = \hat{x}_C(t) + \Delta$, Δ is a constant.
Cyclic	$x_C(t) = \hat{x}_C(t) + e$, e is a random error.
Drift	$x_C(t) = \hat{x}_C(t) + e(t)$, $e(t)$ is a changing error.

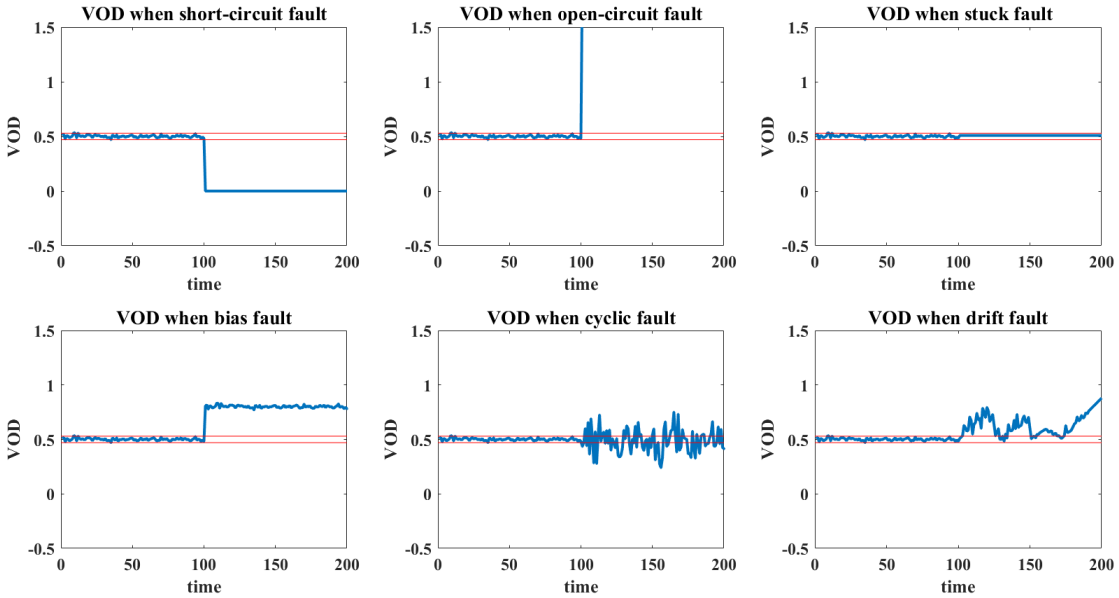

Figure 3.5: Schematic diagrams of VODs when sensor faults occur.

3.2.3.2. Variable of action difference

The variable of action difference (VAD) is the difference in control action by different

participants.

Consider $u(t)$ and $d_u(t)$ are the control action and VAD, respectively.

$$d_u(t) = u_C(t) - u_H(t) \tag{3.2}$$

where $u_C(t)$ is the controller action, and $u_H(t)$ is the human operator's control action. The symbols d and $d(t)$ represent either VOD or VAD in this paper.

As the observation determines the action, suppose

$$u = g(x) \tag{3.3}$$

where g is the function from observation to action.

$$d_u(t) = g\big(x_C(t)\big) - g\big(x_H(t)\big) \tag{3.4}$$

Usually, the human operator works on standby as a monitor or supervisor, who may not take action when the process variable is within the normal range around the setpoint x_0. Human observation can be continuous, while the human control action may not be continuous. To make human action $u_H(t)$ a continuous function, suppose there is a period of no human action from t_0 to t_1, the observation is in the normal range around the setpoint x_0, and the action is the same as the starting point u_0, it also means

$$u_H(t \in [t_0, t_1]) = u_0 = g(x_0) \tag{3.5}$$

Furthermore, the action is usually nonlinear with the observation. The Taylor series expansion is applied to reach a linear approximation around the setpoints x_0 and u_0.

$$u \approx g(x_0) + g'(x_0)(x - x_0) = g'(x_0)x + g(x_0) - g'(x_0)x_0 \tag{3.6}$$

According to the operation rule on a single normal variable, the linear operation of the normal variable is also normally distributed. Based on Equation (3.6) which is a linear

operation, approximatively, VAD follows $d_u \sim N(0, |g'(x_0)\sigma|^2)$. Hence, the judgmental condition of action conflict should be $d_A \notin [-|g'(x_0)\sigma|, |g'(x_0)\sigma|]$.

There are two critical preconditions for an action conflict to occur. The first condition is that the human operator and the controller have equal priority, or the controller has higher priority. Otherwise, the human operator can skip manipulating the controller to take control of the entire process system. The second condition is that an action conflict should be "a confirmed fault" ($x_C \notin [\mu - 3\sigma, \mu + 3\sigma]$, $x_H \notin [\mu - 3\sigma, \mu + 3\sigma]$) at first.

3.2.3.3. Relation between VOD and VAD

In Control Engineering, the function g is usually expressed by a transfer function. The observation can be considered as the action result. For a control system, consider control action $u(t)$ as the input variable, observation $x(t)$ as the state variable, and action result $y(t)$ as the output variable which can be set the same as observation $x(t)$, then the transfer function $G(s)$ is

$$G(s) = \frac{Y(s)}{U(s)} \tag{3.7}$$

Where $Y(s)$, $U(s)$ are the Laplace transforms of $y(t)$ and $u(t)$.

The Laplace transform of VAD is

$$D_u(s) = U_C(s) - U_H(s) \tag{3.8}$$

The Laplace transform of VOD is

$$D_x(s) = Y_C(s) - Y_H(s) = G(s)[U_C(s) - U_H(s)] = G(s)D_u(s) \tag{3.9}$$

It can be concluded that the transfer function from input to output is also applicable from VAD to VOD.

3.2.4. Conflict convergence

The derivative of the conflict function can illustrate the trend of difference. Divergence means the conflict is intensifying and convergence means the conflict is resolving.

For VOD:

$$d_x'(t) = x_C'(t) - x_H'(t) \tag{3.10}$$

For VAD:

$$d_u'(t) = \left[g\left(x_C(t)\right) - g\left(x_H(t)\right)\right]' = g'(x_C(t))x_C'(t) - g'\left(x_H(t)\right)x_H'(t) \tag{3.11}$$

As the process variable may be fluctuating with noise, it is difficult to get the derivatives. The moving average method can be used for smoothing and getting a fitted function. The moving average method creates a series of averages of different subsets of the discrete process data, making the curve smooth and feasible to generate a derivable function. This technique applies to both VOD and VAD. The conditions of conflict convergence are shown in Table 3.2, and the trends are shown in Figure 3.6.

Table 3.2: Conflict convergence conditions.

Difference	**Derivative**	**Conflict convergence**
$d(t) > UCL$	$d'(t) > 0$	Conflict diverging
	$d'(t) < 0$	Conflict converging
	$d'(t) = 0, d'(t-\varepsilon) > 0, d'(t+\varepsilon) < 0$	Stationary points from diverging to converging
	$d'(t) = 0, d'(t-\varepsilon) < 0, d'(t+\varepsilon) > 0$	Stationary points from converging to diverging
	$d'(t) \equiv 0$	Steadiness or unchanged conflict
$d(t) < LCL$	$d'(t) > 0$	Conflict converging
	$d'(t) < 0$	Conflict diverging
	$d'(t) = 0, d'(t-\varepsilon) > 0, d'(t+\varepsilon) < 0$	Stationary points from converging to diverging
	$d'(t) = 0, d'(t-\varepsilon) < 0, d'(t+\varepsilon) > 0$	Stationary points from diverging to converging
	$d'(t) \equiv 0$	Steadiness or unchanged conflict
$LCL < d(t) < UCL$	-	No conflict

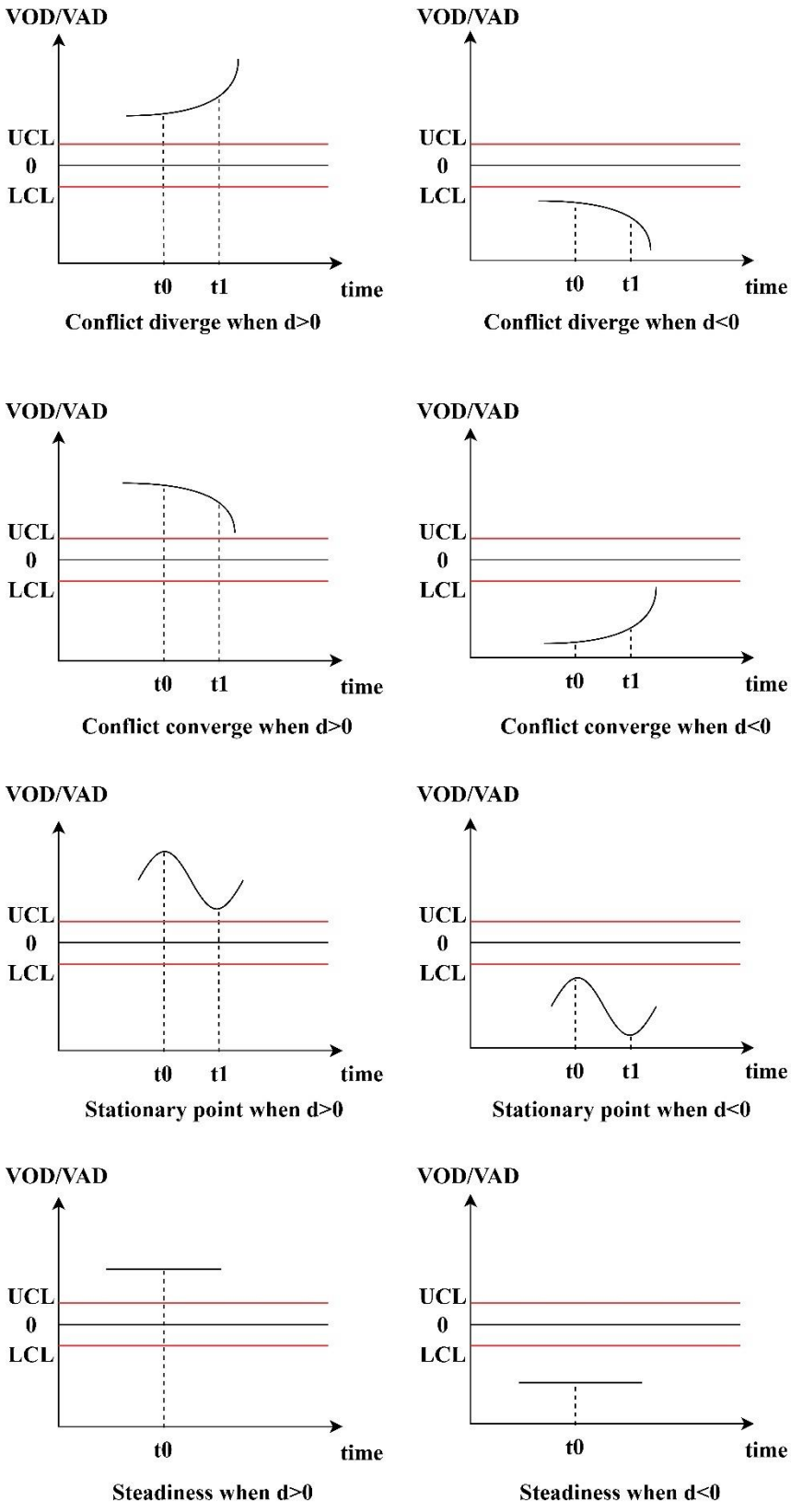

Figure 3.6: Conflict trend.

The comparison of the derivatives also indicates the accelerating or decelerating trends. For example, when $t_0 < t_1$, $d'(t_0) > 0$, $d'(t_1) > 0$, and $d'(t_1) > d'(t_0)$, it can be concluded that the conflict divergence is accelerating, and the contrariwise is decelerating.

3.2.5. Conflict resolution

3.2.5.1. Condition of conflict resolution

For conflict resolution, the task is to reach a stationary point from divergence or steady to convergence by human intervention. The conditions of conflict resolution are shown in Table 3.3, and the trends are shown in Figure 3.7.

Table 3.3: Conflict resolution conditions.

Difference	**Derivative**
$d(t) > UCL$	$d'(t) = 0, d'(t-\varepsilon) \geq 0, d'(t+\varepsilon) < 0$
$d(t) < LCL$	$d'(t) = 0, d'(t-\varepsilon) \leq 0, d'(t+\varepsilon) > 0$

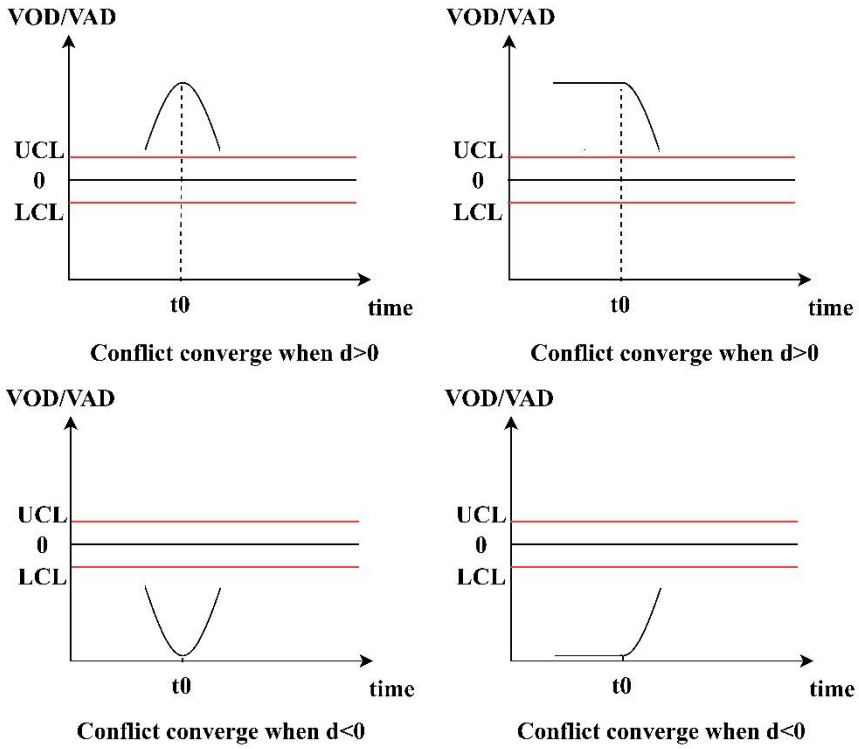

Figure 3.7: Conflict resolution situations.

3.2.5.2. Human intervention for conflict resolution

There is a period for human intervention to resolve the conflict before the stationary point. For observation conflict resolution, if the human intervention is invalid, it may develop into action conflict. For action conflict resolution, if the human intervention is invalid, it may bring risks to the process system. Common human interventions for conflict resolution are shown in Table 3.4. An example is that, currently, some sensors have online flushing devices to exclude the impurity, and the operator could flush the sensing component manually to resolve the fault and conflict.

Table 3.4: Human intervention for conflict resolution.

Fault cause	Human intervention for conflict resolution
Supply problems	Restore power supply Adjust voltage or current
Connection problems	Correct wiring Correct grounding Correct connections and contacts Solve block or breakpoint
Malfunction of the sensor, logic solver, and actuator	Restart Reset Recalibration
Hardware failure	Repair Replace
Environment factor	Eliminate interference
Internal factor	Exclude impurities
Unknown	Automatic recovery for unknown reasons

3.2.6. Conflict probability

Conflict probability is the frequency measure of occurring a conflict. As VOD follows a normal distribution $N(0, \sigma^2)$, observation conflicts locate at the long tail of both sides in the normal distribution, and it also means an observation conflict is a rare event from the holistic perspective. In addition, the cumulative density function (CDF) is

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2\sigma^2}} dt \tag{3.12}$$

Usually, the observation has a maximum value x_{max} and a minimum value x_{min} which are determined by the system and the sensor. For example, the oil level in a tank has a maximum value of the tank height (full of oil) and a minimum value of 0 (no oil). For the range of VOD, it has $d_{max} = x_{max} - x_{min}$ and $d_{min} = x_{min} - x_{max}$.

Referring to the min-max normalization technique (Han et al., 2012), the conflict probability is proposed as

$$P = \begin{cases} \frac{F(LCL)-F(d_x)}{F(LCL)-F(d_{min})}, d_x \le LCL \\ \frac{F(UCL)-F(d_x)}{F(UCL)-F(d_{max})}, d_x \ge UCL \\ 0,\ LCL < d_x < UCL \end{cases} \tag{3.13}$$

As $LCL = -\sigma$, $UCL = \sigma$, $F(-\sigma) = 0.159$, $F(\sigma) = 0.841$, then the conflict probability can be simplified as

$$P = \begin{cases} \frac{0.159-F(d_x)}{0.159-F(d_{min})}, d_x \le -\sigma \\ \frac{0.841-F(d_x)}{0.841-F(d_{max})}, d_x \ge \sigma \\ 0,\ -\sigma < d_x < \sigma \end{cases} \tag{3.14}$$

Figure 3.8 shows an exemplary probability distribution for observation conflict.

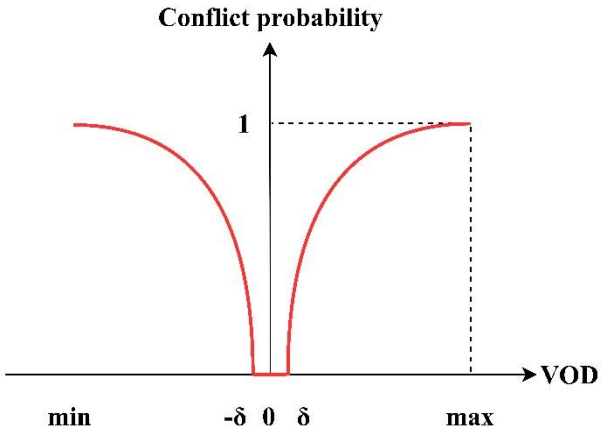

Figure 3.8: The probability distribution of observation conflict.

3.2.7. Conflict risk

Conflict severity is the consequence measure of a conflict. Consequently, conflict risk is the combination of conflict probability and conflict severity. When the conflict diverges, the conflict severity increases significantly. Accordingly, suppose the severity follows the inverse function of a Beta distribution, and it can be expressed as

$$S = \begin{cases} BETA.INV(P, \alpha, \beta), d_x \le -\sigma \ or \ d_x \ge \sigma \\ 0, \qquad -\sigma < d_x < \sigma \end{cases} \tag{3.15}$$

Where $BETA.INV$ is to return the inverse of the beta cumulative probability density

function, and suppose the parameter α is 1, β is 10; P is the probability based on Equation (3.14).

Figure 3.9 shows the example of severity distribution for observation conflict.

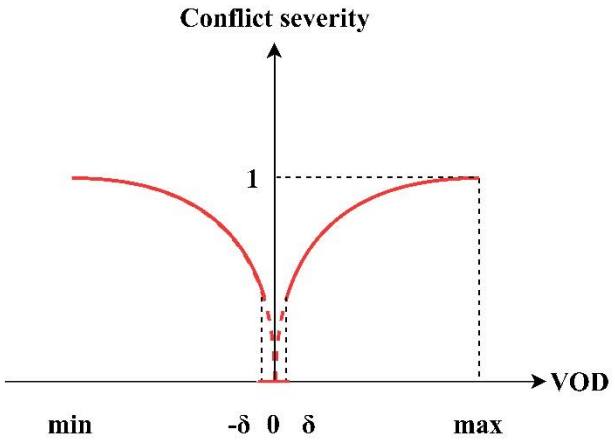

Figure 3.9: The severity distribution of observation conflict.

Consequently, the observation conflict risk is

$$R = P \times S \tag{3.16}$$

Similarly, the above method to calculate probability and risk is also applicable to VAD and the difference is the standard deviation is $|g'(x_0)\sigma|$, instead of σ.

3.3. Application of the methodology

3.3.1. Case description and simulation

The two-phase separator is a common device to separate oil and gas (Figure 3.10). This study set two types of level measurement: a tubular level gauge and a *differential pressure transmitter*. An operator monitored the system by reading the tubular level gauge. The differential pressure transmitter was connected with the level controller and the control valve. The control valve could be adjusted by the controller and the operator

at the same time.

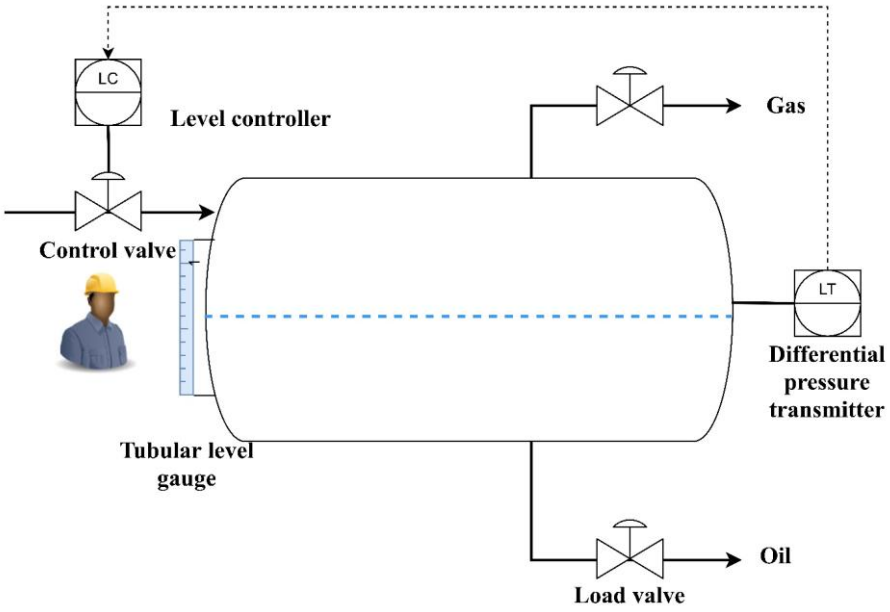

Figure 3.10: Two-phase oil and gas separator.

This study assumed that the crude oil had the same density as water, and the two-phase separator could be considered a conventional water tank level control system. This study adopted detailed assumptions and derivations from a published paper (Zhao & Zhang, 2020). The assumptions were that the cross-sectional area of the tank, setpoint height of oil in the tank, responding valve opening, the height of the tank, cross-sectional area of the pipe, and maximum inflow rate of oil intake were 1 m^2, 0.50 m, 50%, 1 m, 0.005 m^2, and 1 m^3/s, respectively.

For this system, the differential equation was

$$\frac{dV}{dt} = C\frac{dh}{dt} = bu - a\sqrt{h} \tag{3.17}$$

Where V was the volume of oil in the tank, C was the cross-sectional area of the tank, h was the height of oil in the tank, b was a constant related to the flow rate into the tank, u was the valve opening, and a was a constant related to the flow rate out of the

tank.

The transfer function from the input variable u to the output variable h was

$$G(s) = \frac{0.8}{2s^2+s} \tag{3.18}$$

The simulation by MATLAB/Simulink R2021a was proposed in Figure 3.11. A proportional-integral-derivative controller (PID) is simulated as the controller of the automated system, compared with a proportional controller as the human operator. The ramp signals were used to simulate the faults. The variables were listed in Table 3.5.

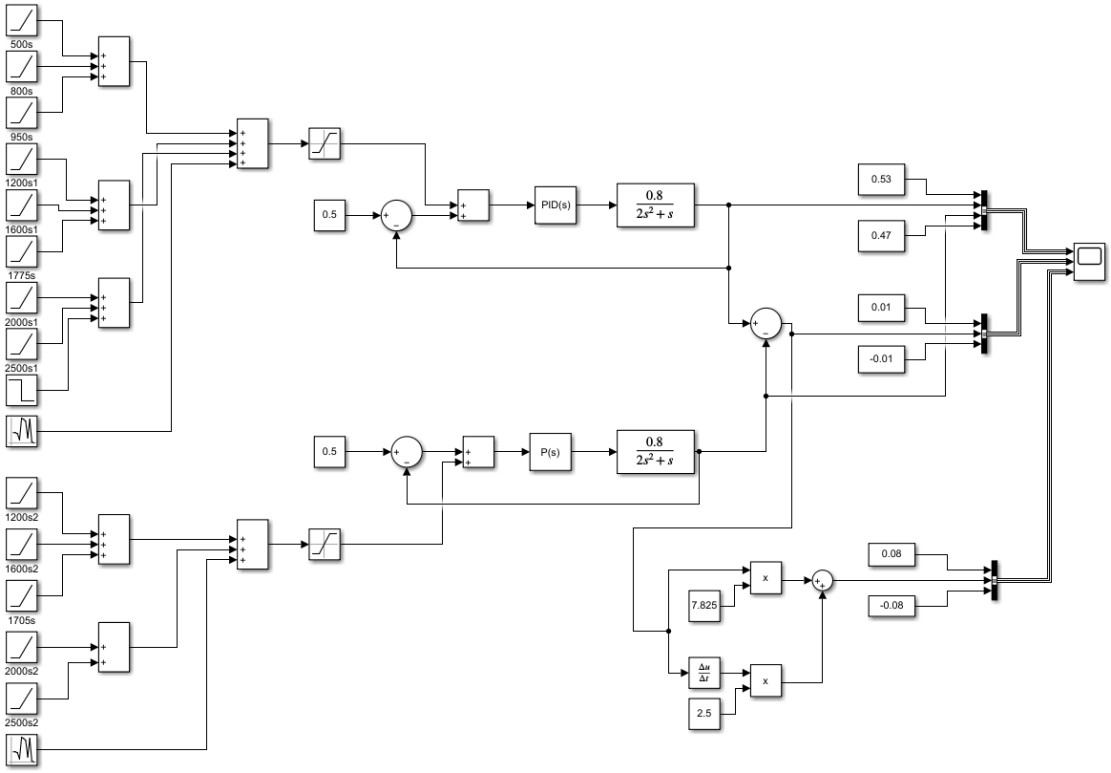

Figure 3.11: Simulink model of observation conflict.

Table 3.5: Variables of the two-phase separator.

Variable	Symbol	Description	Range
Input	$u(t)$	Action: the valve opening	[0,100%]
State	$x(t)$	Observation: the height of oil	[0,1]; $x \sim N(0.5, 0.01^2)$
Output	$y(t)$	Observation: the height of oil	[0,1]; $x \sim N(0.5, 0.01^2)$
VOD	$d_x(t)$	Observation difference	[-1,1]
VAD	$d_u(t)$	Action difference	[-100%,100%]

The simulation was conducted from 0 s to 3000 s and 3 faults were presented based on

Scenario 2, 4, 6 in Section 3.2.2 respectively. The observations were recorded in Figure 3.12.

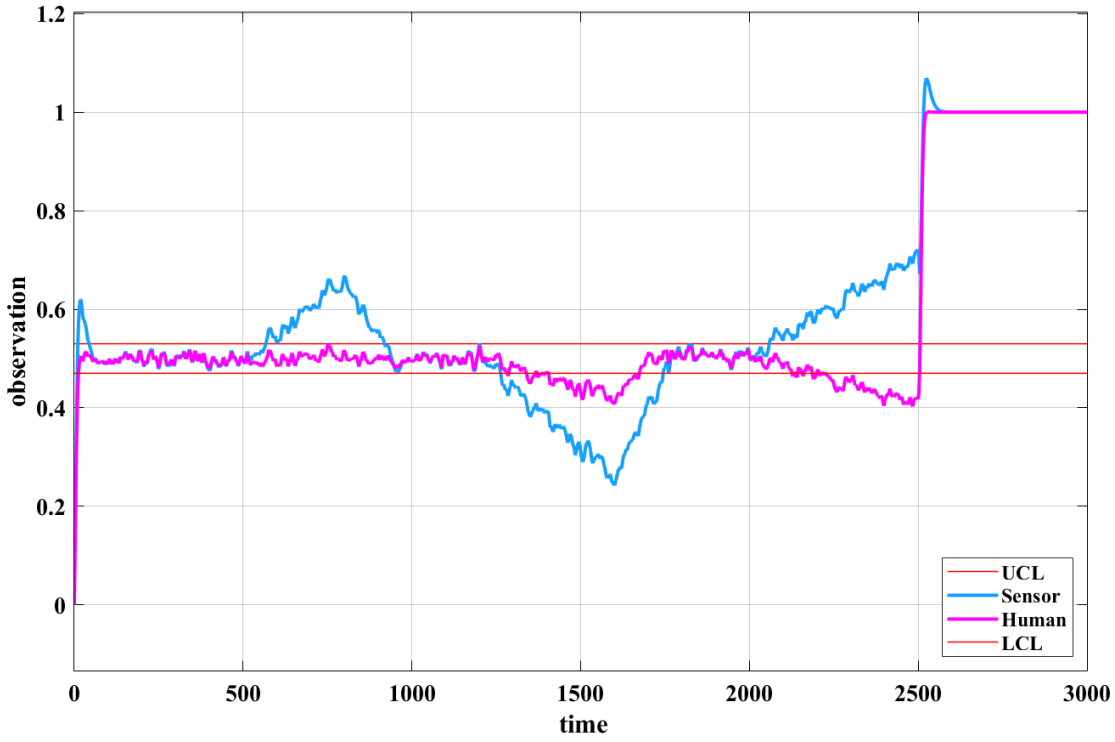

Figure 3.12: The observations of the oil level.

3.3.2. Conflict evolution

3.3.2.1. Conflict 1: observation conflict not action conflict

This was a simulation of Scenario 2 in Section 3.2.2. At 500 s, the sensor observed the oil level increasing by the transmitter. At 600 s, the sensor detected the oil level out of limit and the controller generated a fault alarm. However, the human operator observed normal by the tubular level gauge. An observation conflict occurred at 600 s. The operator supposed that the sand in the crude oil may have contaminated the sensing element of the transmitter. At 800 s, the operator activated the flanged flushing device of the differential pressure transmitter. The oil level started to drop. At 900 s, the sensor

observation fell to normal. The real oil level was in the normal range all the time. Thus, it was not a confirmed fault by the operator. In addition, the operator did not take any control action on the control valve, therefore, it was not an action conflict. This fault was a malfunction of the transmitter due to the sand impurity in crude and did not activate the controller action.

3.3.2.2. Conflict 2: action conflict with resolution

This was a simulation of Scenario 4 in Section 3.2.2. At 1200 s, the sensor observed the oil level decreasing by the transmitter. At 1300 s, the sensor detected the oil level out of limit and the controller generated a fault alarm. The human operator also observed the oil level decreasing by the ruler; however, the two observations were significantly different. An observation conflict occurred at 1300 s. Then the operator checked but did not find any failure. The controller increased the valve opening, and at the same time, the operator tried to stop the controller from increasing too much. An action conflict occurred at 1300 s. At 1600 s, the operator found the transmitter indicating wrong numbers and then reset the transmitter. The oil level started to increase. At 1700 s, the oil level was back to normal. This fault was an indicating error of the transmitter and caused observation conflict and action conflict.

3.3.2.3. Conflict 3: action conflict without resolution

This was a simulation of Scenario 6 in Section 3.2.2. At 2000 s, the sensor observed the oil level increasing by the transmitter. At 2060 s, the sensor observed the oil level out of the limit and generated a fault alarm. However, the human operator observed the oil

level decreasing by the ruler. An observation conflict occurred at 2060 s. Then the operator checked but did not find any failure. The controller decreased the valve opening, and at the same time, the operator increased the valve opening. An action conflict occurred at 2060 s. At 2500 s, the valve experienced a mechanical failure due to the action conflict on it. The valve opening was increased to the maximum. At 2520 s, the tank overflowed. This conflict was unsolved and led to a spill accident.

3.3.3. Conflict variables

The observations, VOD, and VAD were shown in Figure 3.13.

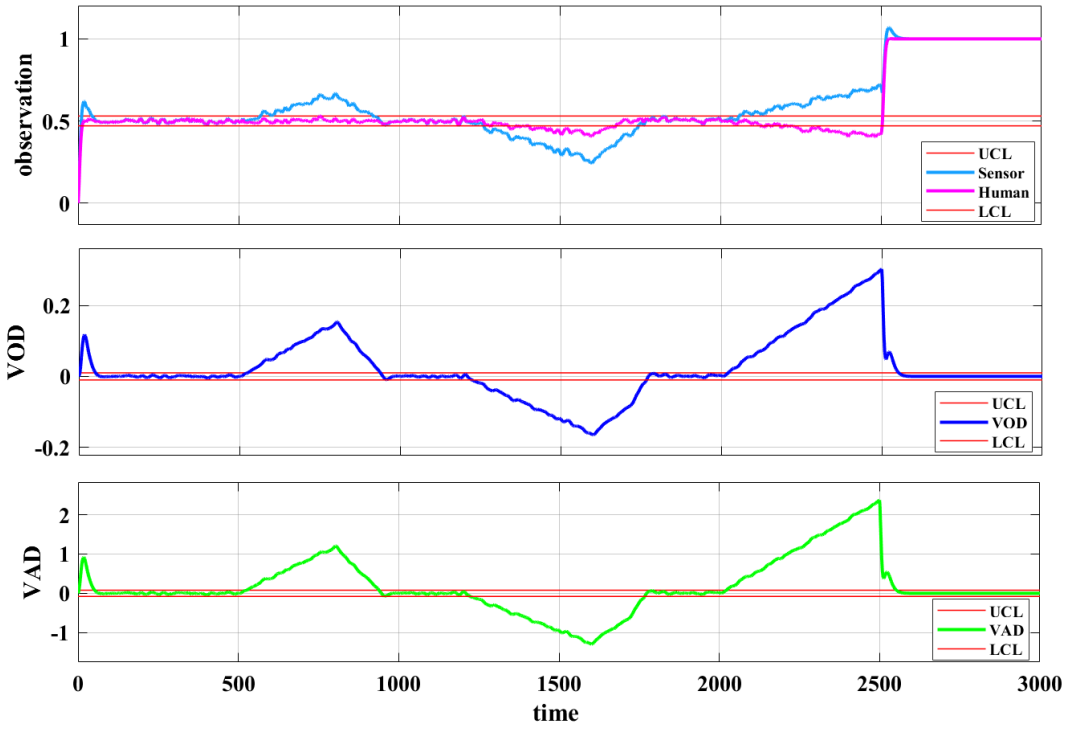

Figure 3.13: Observations, VOD, and VAD.

3.3.4. Conflict convergence

A piecewise linear fit was performed on the VOD to verify the conflict convergence. The slope (the derivative of the piecewise function) could indicate the function was

increasing or decreasing, in other words, it could indicate VOD was diverging or converging. For example, in 500-800 s, use the fit function of MATLAB and get the slope 0.00051, which means $d_x' = 0.00051 > 0$; in this period, $d_x > UCL$, according to the condition of conflict convergence (Table 3.2), it can be concluded that the conflict is diverging in 500-800 s. This method was applied to VOD and VAD to get the convergence situations (Table 3.6).

Table 3.6: Conflict convergence.

t	d_x	d_x'	Observation conflict	d_u	d_u'	Action conflict
600-799 s	>UCL	>0	Conflict diverging	-	-	-
800 s	>UCL	0	Stationary point			
801-900 s	>UCL	<0	Conflict converging	-	-	-
1300-1599 s	<LCL	<0	Conflict diverging	<LCL*	<0	Conflict diverging
1600 s	<LCL	0	Stationary point	<LCL*	0	Stationary point
1601-1700 s	<LCL	>0	Conflict converging	<LCL*	>0	Conflict converging
2060-2499 s	>UCL	>0	Conflict diverging	>UCL*	>0	Conflict diverging
2500 s	>UCL	0	Stationary point	>UCL*	0	Stationary point
2501-2520 s	>UCL	$-\infty$	-	>UCL*	$-\infty$	-

Here the symbol * meant the limits of VOD and VAD were different.

3.3.5. Conflict resolution

From the above analysis, the situations at 800 s and 1600 s met the condition of conflict resolution. At 800 s, the operator removed the interference. At 1600 s, the operator reset the transmitter. The operator conducted effective interventions to resolve the conflicts. For 2500 s, the operator did not intervene, only worked on the control valve. Furthermore, the action conflict resulted in the valve failure which caused the valve

opening to the maximum, and the oil filled the tank and overflowed. At this moment (2500 s-2520 s), the VOD and VOD were shrunk to 0. However, it was meaningless to indicate the conflict convergence.

3.3.6. Conflict probability, severity, and risk

Though the final risk was caused by VAD, VOD and VAD kept the same trend in this case. Here took VOD as an example to calculate the probability and risk, and the period 2000 s-2500 s was selected. At 2030 s, the VOD exceeded the limit. At 2060 s, the controller generated an alarm. At 2500 s, the controller experienced a failure due to action conflict. The selected object timepoints and data were shown in Table 3.7.

According to the proposed method, Equation (3.12) and Equation (3.14) were used to calculate the observation conflict probability at each timepoint. Equation (3.15) was used to get the severity. Then the observation conflict risk at each timepoint could be estimated by Equation (3.16). The probabilities, severities, and risks at selected timepoints were also shown in Table 3.7.

Table 3.7: Values in selected timepoints.

t	d_x	x_C	Probability	Severity	Risk
2030 s	0.02	0.52	0.70	0.12	0.09
2060 s	0.04	0.54	0.99	0.50	0.50
2080 s	0.05	0.56	1.00	0.77	0.77
2500 s	0.30	0.72	1.00	1.00	1.00

The risks and severities raised significantly and rapidly as conflict diverged. At 2080 s, the conflict probability had been approaching 100% and the risk became certain with its responding severity. It explained that the conflict can cause severe consequences in a quite short period, and there was not adequate time left for the human operator to

respond effectively. This has been seen in Boeing 737 Max accidents and phantom braking scenarios.

3.4. Discussion

This study assumes that there are two observers, human and automated system, for the same process variable, and the proposed method is to assume the expectation of human observation is consistent with sensor observation. Practically, human observation has a larger inaccuracy problem. Yet it is not significant compared with the fault data, not affecting the calculation of this study.

For the judgment limit, in this study, the VOD limit is set as $\pm\sigma$, which may be relatively strict. As the process value fluctuates $\mu \pm 3\sigma$ which is considered the normal range, it encounters that sometimes the VOD is out of the limit while the process value is still normal, not triggering a fault alarm. Fortunately, it indicates that conflicts are more sensitive than faults, and it is more valuable to predict conflicts.

This study uses a univariate method which makes the whole process easy to illustrate. Future research should consider the scenarios of multidimensional variables in real complex systems. Dimensionality reduction methods, for example, principal component analysis (PCA), may be applied. Challenges associated with the human-machine teaming in the multi-agent system (Canonico, 2019), and the multi-input multi-output problems (Ahmed, 2016; Ahmed & Imtiaz, 2015) should be investigated.

As faults are the symptom of conflicts, the mature techniques of fault diagnosis and resolution could be considered in conflict research. The model predictive control (MPC),

linear quadratic regulator (LQR), and data-driven control may smooth or offset the impact of the fault and even conflict. The non-linear MPC, explicit MPC, and robust MPC (Pistikopoulos, 2009), may provide a better optimization solution compared with PID. Therefore, robust design, fault-tolerant design, and data-driven control might contribute to the inhibition and resolution of conflict. In the meantime, the conflict analysis might be utilized to consider the constraints and boundaries of the robust control design. However, it should be noted that data-driven or AI-based control may show lower understandability and interpretability with their black box nature (Ahmed, 2021).

In addition, another source of conflict is human error and intentional human action, which are often characterized as security issues that may trigger more unpredictable, unreasonable, and severe conflicts. The real scenarios to apply this conflict methodology are to be excavated and anticipated before a catastrophe shows up in the process industry due to conflicts.

For conflict resolution, some techniques of human intervention have been listed and these are the direct measures to solve the sensor fault first which is the source of the conflict. Moreover, traditional methods of improving sensor reliability, such as redundant design and risk-based maintenance, are still proactive. Once the sensor fault is confirmed, the higher priority of the human operator should be unambiguous, and it needs to be convinced that the system could be switched to manual mode. Respectively, straightforward procedures on how to solve the conflict should be delivered to the

human operator.

For conflict risk mitigation, full automation is the ultimate risk elimination, yet the reality is that the human-automated system collaboration will still exist for a long time. Therefore, in the design phase of the human-automated systems, the conflict analysis should be considered and stressed, which is further thinking beyond reliability issues. In the meantime, the human-centered design should balance human reliability and conflict resolution by human intervention.

3.5. Conclusions

This study systematically illustrated the concept and definition of human-automated system conflict in the process industry, presented the mathematical expression of observation conflict and action conflict, discovered the convergence and resolution conditions of conflicts, and applied a case study on a classic model of the two-phase separator. Different from previous fault diagnosis research, conflict is another deeper and more implicit phenomenon that brings risks more rapidly and severely. Conflicts are highly associated with faults and failures, furthermore, faults are the symptom of conflicts, and failures are often correlated with conflicts.

The automated systems cannot work alone without the supervision of humans, as they still cannot give value judgment, even though they are of higher reliability than humans. The human-automated system conflict may be triggered due to resistance to human participation. However, once the automated system fails, the consequence cannot be corrected by itself. Such conflicts can deteriorate the consequences and are more

difficult to deal with than human error. Therefore, human-centered design is required when the automated system is applied, and this would contribute to reducing the occurrences of human-automated system conflict and approaching the future of human-automated system collaboration, even human-AI collaboration.

Chapter 4: Conflict Due to Cyberattack

Preface

A version of this chapter has been published in *Computers & Chemical Engineering*. I am the primary author, along with the co-authors, Drs. Faisal Khan, Salim Ahmed, Syed Imtiaz, and Stratos Pistikopoulos. I developed the conceptual framework for the methodology and carried out the literature review and case study. I prepared the first draft of the manuscript and subsequently revised the manuscript based on the co-authors' and peer review feedback. Co-author Dr. Faisal Khan helped develop the concept, verify the methodology, review, and revise the manuscript. Co- authors Drs. Salim Ahmed, Syed Imtiaz, and Stratos Pistikopoulos provided support in implementing the concept and verifying the methodology. The co-authors provided fundamental assistance in validating, reviewing, and correcting the methodology, case study, and results. The co-authors also contributed to the review and revision of the manuscript.

Reference: Wen, H., Khan, F., Ahmed, S., Imtiaz, S., & Pistikopoulos, S. (2023). Risk assessment of human-automation conflict under cyberattacks in process systems. *Computers & Chemical Engineering*, 172, 108175. https://doi.org/https://doi.org/10.1016/j.compchemeng.2023.108175

Abstract

Human-automation conflict is a frontier subject to be vigilant against, and it may become even more elusive and confusing under cyberattacks. Conflicts due to cyberattacks are often beyond the expertise of the human operator, making the

resolution much more difficult. Therefore, this study transforms common attacks into understandable representations with process variables, explores human-automation conflict under five generalized attacks, e.g., false data injection, denial of service, etc., and explains the conflict with a proposed cooperative Pareto paradigm based on game theory, then applies on a Continuous Stirred Tank Reactor. The results show that cyberattacks could cause conflicts significantly. It also highlights that attack strength determines the strength of action conflict. The control actions could buffer the finite impact of attacks within a limited range. The conflict risk could be applied to distinguish a fault and an attack, and measures could be taken accordingly.

Keywords: Conflict, cyberattack, game theory, human-automation conflict.

4.1. Introduction

As systems become digitized and automated, human involvement is reduced or dispensable; even worse, machines could exclude human intervention. This situation is regarded as human-automation conflict (Wen et al., 2022) or human-machine conflict, which has been demonstrated when sensor faults occur, for example, in the Boeing 737 Max accidents (DeFazio & Larsen, 2020). In addition, cyberattacks may also be supposed to trigger such conflicts (Figure 4.1), for example, false data injection (FDI) on sensor, which is equivalent to sensor faults in terms of consequences.

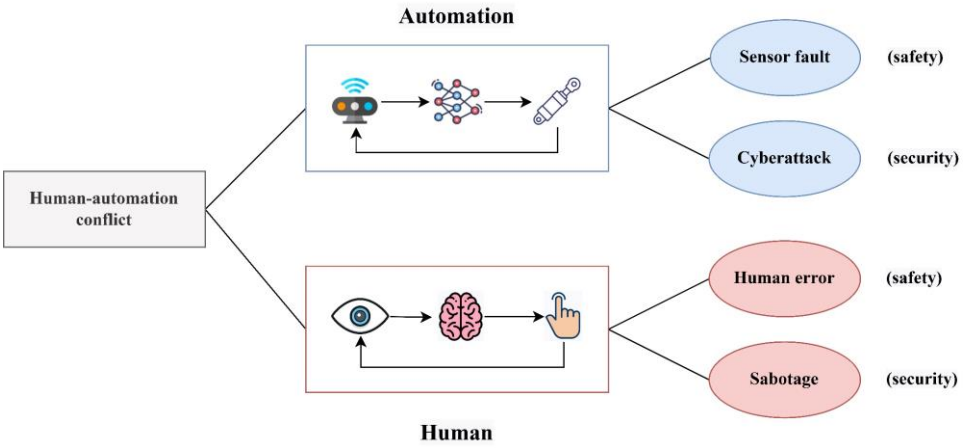

Figure 4.1: Human-automation conflict.

There have been numerous notable cyber incidents that triggered human-automation conflicts. One example is Stuxnet, which is deemed to be the first malware targeting industrial control systems (ICS) (Chen, 2010; Kushner, 2013). The malware caused the centrifuges to overspin, leading to wear and tear; though the operator was able to detect the abnormality, it was an arduous task to resolve it. Another example is that a hacker accessed a water plant's human-machine interface (HMI) and tampered with sodium hydroxide's level from 100 ppm to 11,100 ppm (Campo-Flores, 2021). The operator witnessed a peculiar phenomenon in that the value was modified repeatedly after correction.

The above examples demonstrate that cyberattacks may manifest the phenomenon of human-automation conflict. Based on industrial reports, cyber threats to process systems have been growing tremendously (European Union Agency for Cybersecurity [ENISA], 2022; Kaspersky, 2022). However, few studies have drawn attention to the fact that cyberattacks result in human-automation conflict. From the analysis of historical cyber incidents (Hemsley et al., 2018; Iaiani, Tugnoli, Bonvicini, et al., 2021b;

Iaiani, Tugnoli, Macini, et al., 2021), two categories of cyberattacks can prompt human-automation conflict. One is impairment to network transmission performance, such as Denial of Service (DoS) and time delay, especially on programmable logic controllers (PLC) (Ylmaz et al., 2018). The other is the manipulation of the process variables and parameters, for example, setpoint modification, FDI on sensor or actuator (Liang et al., 2017).

Traditionally, cyberattacks have been explored separately in the cyber domain or regarded as faults in process system engineering rather than conflicts. More focus stresses on intrusion detection (Giraldo et al., 2018; Narasimhan et al., 2022) and fault diagnosis (Syfert et al., 2022). Some literature has presented progressive outputs on the attack resistance of model predictive control (MPC) and other advanced control (Arauz et al., 2022; Rangan et al., 2022), with applications on classic process models, as an example, Continuous Stirred Tank Reactor (CSTR) (Durand & Wegener, 2020; Z. Wu et al., 2018).

Moreover, the literature displays that the approaches proposed by process engineers and IT engineers are entirely dissimilar as they apply different technical languages. The operators cannot recognize cyberattacks with process engineering expertise. This heightens the possibility of human-automation conflicts. Furthermore, such conflicts are fully manifested in the struggle for control authority, which is a typical competitive game between the operator and the hacker. Therefore, researchers have attempted to simulate attacks based on game theory. Consequently, various game paradigms have

been proposed (Liu et al., 2021; Nikmehr & Moghadam, 2019; Shen et al., 2021). This inspires the path to interpret human-automation conflict.

In general, the links between cyber security and process safety have not been well established, and the paths to how these attack techniques impair and disrupt field devices have not been established. On this basis, this study involves two issues: conflict and game. Hence, an urgent need is to assist operators in identifying attacks and distinguishing them from faults, then resolving conflicts calmly. Therefore, this study attempts to explore the human-automation conflict under cyberattack, illustrate the conflict with game theory, establish a risk model to distinguish the attack and fault, and apply it to a classic CSTR. This study attempts to answer the following research questions:

i. How to distinguish an attack from a fault?
ii. How to represent a potential cyberattack with process variables?
iii. How to identify conflict under attack?
iv. How to illustrate the conflict with the game paradigm?
v. How to assess the conflict risk under attack and guide risk management?

The present study introduces several novel concepts and mathematical formulations, which include: i) extension of human-automation conflict under cyberattack further than sensor fault; ii) attack representation with process variables for easier understanding to operators; iii) mathematical formulation of conflicts with cooperative Pareto paradigm; iv) illustration of response to the attack through robust control with

bounded or unbounded attack strength, e.g., MPC; and v) risk model of human-automation conflict under attack.

This article is organized as follows: Section 4.2 presents the methodology; Section 4.3 describes the application to a CSTR; Section 4.4 is the discussion and critical observations, highlighting the assumptions and the importance of this work; Section 4.5 is the concluding remarks and future work.

4.2. Methodology

4.2.1. General description

The methodology employed in the present study is shown in Figure 4.2, with five significant steps. The more detailed illustration refers to the following subsections.

Step 1: Transform and represent attacks with process variables.

Step 2: Identify conflicts under attack.

Step 3: Explain the conflict with the game paradigm.

Step 4: Assess the conflict probability.

Step 5: Quantify the conflict severity and risk.

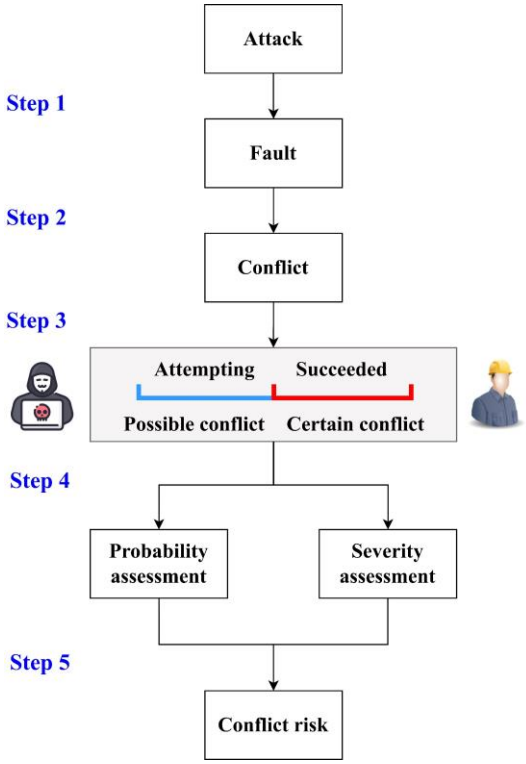

Figure 4.2: The research methodology.

4.2.2. Transform and represent attacks

Human operators are the primary responder to abnormal situations, with inadequate expertise to rebel against cyberattacks. This gap should be filled by transforming cyberattacks into understandable process variables and parameters. Based on the historical cyber incidents (Hemsley et al., 2018), review papers (Iaiani, Tugnoli, Bonvicini, et al., 2021b; Iaiani, Tugnoli, Macini, et al., 2021), reports (Bundesamt für Sicherheit in der Informationstechnik [BSI], 2022), and brainstorming, the general cyberattack methods, techniques, and consequences are identified in Table 4.1. The cyberattack methods can be independent or combined. The specified cyberattack methods can be shown on a closed-loop control diagram (Figure 4.3).

Table 4.1: Cyberattack methods on a closed-loop control system.

Attack method	Technique	Consequence
FDI on sensor	Manipulate the reading of the sensor, or manipulate the data from the sensor to the controller	Faulty sensor data
Setpoint modification	Modify the setpoint	Undesired steady state
FDI on actuator	Manipulate the input value	Undesired input
DoS	Crash the computing or communication ability of the server and other devices	No network service
Time delay	Manipulate the program or coding, change network configuration, or block the data in transmission	Network service degradation

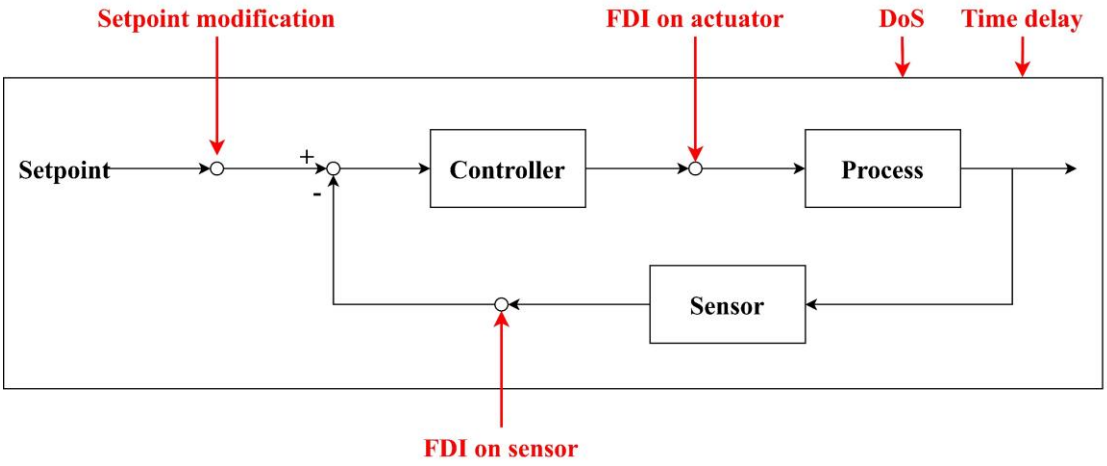

Figure 4.3: Cyberattack methods on a closed-loop control system.

For a general linear time-invariant system, it has

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t) \\ y(t) = Cx(t) + Du(t) \end{cases} \tag{4.1}$$

$$\frac{Y(s)}{R(s)} = \frac{G(s)H(s)}{1+G(s)H(s)} \tag{4.2}$$

Where in the time domain, x is the state variable or the observation without attack, which usually follows a normal distribution $N(\mu, \sigma)$; u is the input variable or the

action without attack; y is the output variable; A, B, C, and D are the state, input, output, feedthrough matrices respectively. In s domain, Y is the output function, R is the reference function, G is the transfer function of the controller, H is the transfer function of the process.

It can be derived that each type of attack could be represented in the time domain and s domain (Table 4.2), where r is setpoint, $\tilde{r}$ is the modified setpoint, $\tilde{u}$ is the input in hacked status, $\tilde{x}$ is observation in hacked status, and the tilde donates the hacked status; w is the attack vector, w_1 is FDI on sensor, w_2 is setpoint modification, w_3 is FDI on actuator, w_4 is DoS, w_5 is a time delay attack, and τ is delayed time; U is input variable or action in the s domain, $\widetilde{U}$ is hacked input variable or action in the s domain, W is attack in the s domain.

Table 4.2: Representation of attack.

Attack method	**Time domain**	**s domain**
FDI on sensor	$\tilde{x}(t) = x(t) + w_1(t)$	$\widetilde{U}(s) = U(s) + R(s)G(s)W_1(s)$
Setpoint modification	$\tilde{x}(t) = x(t) + \tilde{r} - r$	$\widetilde{U}(s) = U(s) + R(s)G(s)W_2(s)$
FDI on actuator	$\tilde{u}(t) = u(t) + w_3(t)$	$\widetilde{U}(s) = U(s) + W_3(s)$
DoS	$\tilde{u}(t) = 0$	$\widetilde{U}(s) = U(s) + [-U(s)]$
Time delay	$\tilde{u}(t) = \begin{cases} 0, & t \in [0, \tau] \\ u(t), & t \in (\tau, \infty) \end{cases}$	$\widetilde{U}(s) = \begin{cases} 0, & t \in [0, \tau] \\ U(s), & t \in (\tau, \infty) \end{cases}$

From the representations in s domain, the control action under attack could be considered as the sum of the control action without attack and the attack vector, and the general form is

$$\widetilde{U}(s) = U(s) + V(s) \tag{4.3}$$

Where V is the generalized attack in the s domain. And then, it can be transferred to

the time domain, which is

$$\tilde{u}(t) = u(t) + v(t) \tag{4.4}$$

Where v is the generalized attack.

It means all cyberattacks will eventually embody the actuator action, forcing the control system into the wrong action. Also, it will reflect on the sensor observation at the next time step.

4.2.3. Identify conflicts under attack

For human-automation conflict, (Wen et al., 2022) have defined the definition and explored the convergence and resolution conditions. The propositions of observation conflict and action conflict are rewritten below.

$$VOD = x(t) - \hat{x}(t) \tag{4.5}$$

$$VAD = u(t) - \hat{u}(t) \tag{4.6}$$

Where VOD is the variable of observation difference to measure observation conflict; VAD is the variable of action difference to measure action conflict; $\hat{x}(t)$ is the human observation or expectation; $\hat{u}(t)$ is the human action or human input.

From the definition of VAD, without attack, VAD is 0, then it has $u(t) = \hat{u}(t)$; under attack, it has

$$VAD = \tilde{u}(t) - \hat{u}(t) = u(t) + v(t) - \hat{u}(t) = v(t) \tag{4.7}$$

It means once an attack triggers an action conflict, the strength of the attack is just the action difference. All action conflicts will reflect on the observations at the next time step, and corresponding observation conflicts will emerge. The rule to identify

observation conflict is that VOD is beyond $\pm\sigma$; consequently, if the human intervention cannot resolve the observation conflict, an action conflict occurs.

4.2.4. Explain conflict with game paradigm

Sensor faults usually lead to observation conflict, while cyberattacks cause action conflict straightforwardly. An action conflict could be transformed into a game between the hacker and the operator (Figure 4.4), striving for the dominance of the control system. Specifically, the game only exists in the attempting phase before the control system is breached. Thenceforth, an action conflict is inevitable.

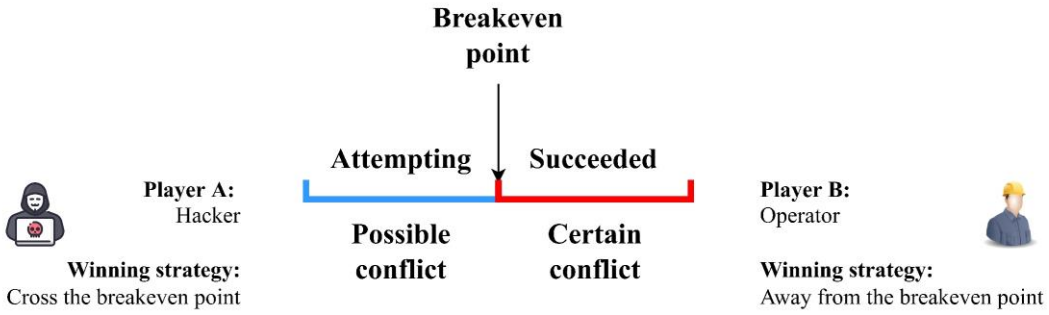

Figure 4.4: The game between hacker and operator.

The action conflict is a zero-sum and non-cooperative game; hence, it does not exist as a pure strategy Nash equilibrium to achieve a double win. However, there should be an equilibrium that regards the breakeven point with minimum damage. Therefore, this study transforms the uncooperative game into a cooperative game and develops the Pareto optimal strategy for the operator, which is called the cooperative Pareto paradigm. Model predictive control, robust control, or other advanced control may perform satisfactorily to bounded disturbance. However, cyberattacks usually initiate unbounded disturbances. It is foreseeable that they may exhibit evident resistance when

the disruption is negligible. This study takes the MPC as an example to illustrate the evolution process of this game and the resistance (Figure 4.5).

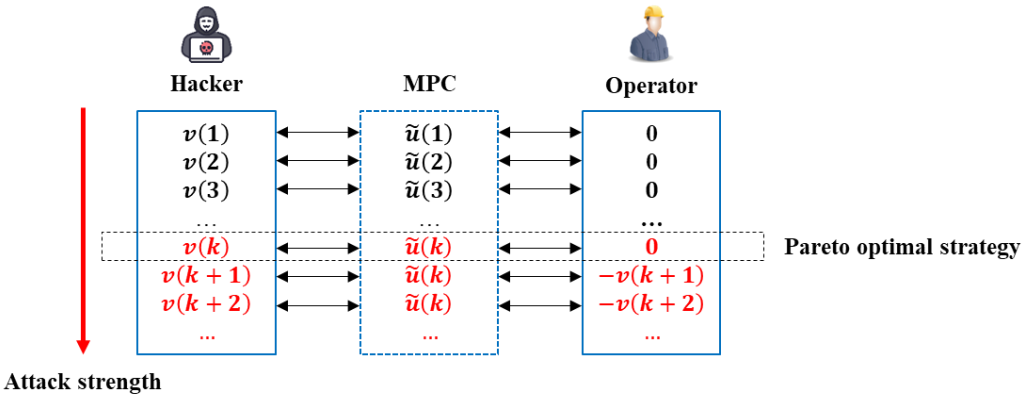

Figure 4.5: Cooperative Pareto paradigm between hacker and operator.

The attack succeeds if the attack's strength forces the MPC to exceed its input constraints. The boundaries of the MPC are exactly the constraints of the game. Also, the input constraints just correspond to the Pareto optimal strategy. Henceforth, the operator must take the exact opposite action to counteract the hacker's attack, which constitutes an action conflict.

MPC has input constraints

$$u_{min} \leq u \leq u_{max} \tag{4.8}$$

The cost function J of MPC with quadratic programming is

$$J = u(k)^T M u(k) + 2x(k)^T O^T u(k) + x(k)^T Q x(k) \tag{4.9}$$

Where M, O, and Q are the achievable matrices in MPC control.

From the safety perspective with the cooperative Pareto paradigm, to maximize the payoff function of the human operator is to minimize the cost function with the optimal $\tilde{u}(k)$.

$$min\ J = min\ [\tilde{u}(k)^T M\tilde{u}(k) + 2x(k)O^T\tilde{u}(k) + x(k)^T Qx(k)] \quad (4.10)$$

Where $\tilde{u}(k) = u(k) + v(k)$.

When $v(k)$ is within the upper and lower constraint, the attack may be directly offset by MPC and it does not require human control action. However, it may also show abnormal action, and it is possible to occur an action conflict. When $v(k)$ exceeds the constraints, that is the failure state, and an action conflict incontrovertibly occurs.

4.2.5. Assess conflict probability

Suppose the system's input variable is the independent variable; the conflict probability and severity can be considered the dependent variables. Therefore, the probability and severity models could be constructed. Based on the definition of VAD and the derivation of the input variable under attack, the more the input deviates from its setpoint, the more the conflict probability approaches 1. Once the attack succeeds, the probability of action conflict is 1. Thus, the probability of action conflict could be proposed (Figure 4.6).

$$P = \begin{cases} 1, u \le \bar{u} - a, u \ge \bar{u} + a \\ BETA.INV(\frac{|u-\bar{u}|}{a}, \alpha, \beta), \bar{u} - a < u < \bar{u} + a \end{cases} \quad (4.11)$$

Where $\bar{u}$ is the input at the setpoint, supposing the input is symmetric; a is the half range of the adjustable input, $a > 0$, u_{max} is $\bar{u} + a$, u_{min} is $\bar{u} - a$; $BETA.INV$ is to return the inverse of the beta cumulative distribution function. The constraint may be asymmetric, here supposing the symmetric situation to simplify the problem.

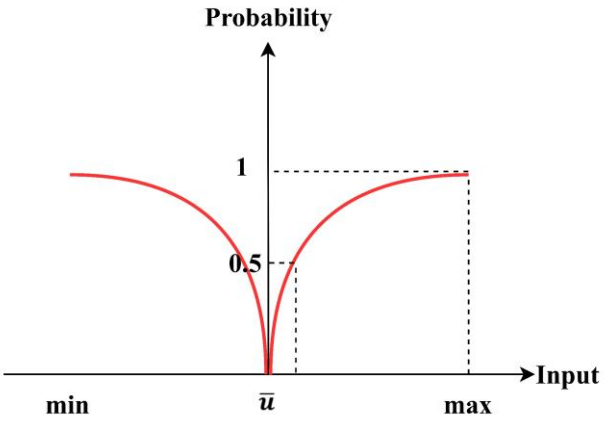

Figure 4.6: Probability of action conflict.

The conflict probability could also contribute to identifying whether it is a fault or an attack. Suppose $P(attack|conflict) + P(fault|conflict) = 1$,

- when $P(attack|conflcit) < P(fault|conflict)$, it is more likely a fault and fault resolution methods could be applied to solve the conflict.
- when $P(attack|conflict) \geq P(fault|conflict)$, it is necessary to consider the system under attack, and the decision should switch to IT solutions, for example, disconnecting the network.

4.2.6. Quantify conflict severity and risk

The severity of action conflict could be proposed as (Figure 4.7)

$$S = (u - \bar{u})^2 \tag{4.12}$$

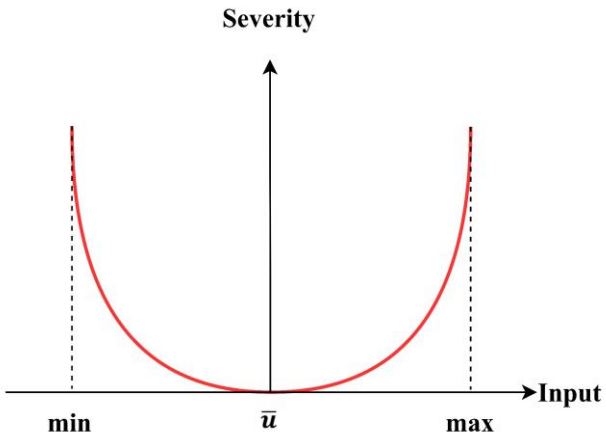

Figure 4.7: Severity of action conflict.

Consequently, the conflict risk is

$$Risk = P \times S \tag{4.13}$$

In most cases, it is strenuous for the operator to anticipate that this is an attack, and the operator tends to resort to traditional fault resolution. Usually, the risk can be graded and set the bar to switch the decision strategy. In this study, half of the maximum risk could be considered a turning point to switch from fault resolution to attack resolution. A comprehensive decision strategy could be summarized in Table 4.3 and evaluated in real application scenarios.

Table 4.3: Decision strategy.

Decision	**Risk**	**Probability**
Fault resolution	$[0,0.5R_{max})$	[0, 0.5)
Attack resolution	$[0.5R_{max},\ R_{max}]$	[0.5, 1]

Where R_{max} is the maximum risk at the upper and lower limit of input.

4.3. Application on CSTR

4.3.1. CSTR description

This study used the classical CSTR model with MPC control by MATLAB & Simulink

2022a (Bemporad et al., 2021). The variables, parameters, and conditions were rewritten in Table 4.4. This study applied a nonlinear CSTR model and then linearized it using MATLAB code. Therefore, the linearized model of CSTR could follow the format of Equation (4.1). In addition, this research utilized the built-in modules and blocks of Simulink and applied MPC Designer to adjust the values of variables and parameters.

Table 4.4: CSTR variable and parameter.

Variable	Description	Initial	Equilibrium	Range
u_1	C_{Af}, the concentration of the reagent in the inlet feed stream, measured in $kmol/m^3$	10	10	-
u_2	T_f, the temperature of the inlet feed stream, measured in K	300	300	$N(300,1)$
u_3	T_c, the temperature of the jacket coolant, measured in K	292	299	[276, 322]
x_1	C_A, the concentration of the reagent in the reactor, measured in $kmol/m^3$	8.5	2	[0, 10]
x_2	T, the temperature in the reactor, measured in K	311	373	[310, 390]

4.3.2. Transform and represent attacks

Based on the CSTR model and cyberattack methods, the hacker may attack the following components (Figure 4.8). Suppose there were two sets of sensing systems for the operator and the controller separately. The operator read the digital value on HMI and acted by digital input to adjust the coolant temperature.

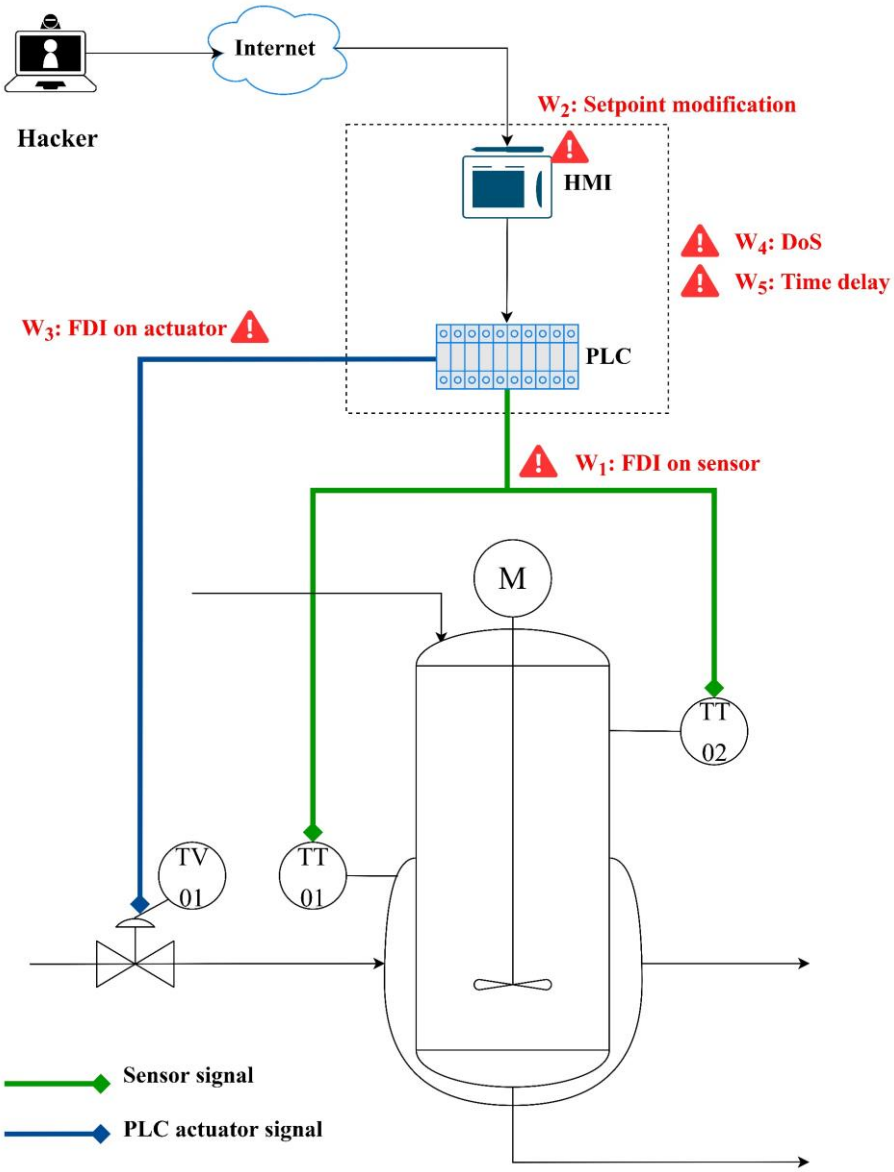

Figure 4.8: Possible cyberattacks on CSTR.

The simulation model was established and shown in Figure 4.9. The cyberattacks were simulated with ramp functions and random functions. The human observation and action were supposed to be consistent with MPC without attack. The difference comparisons of output (tank concentration) and input (coolant temperature) were considered observation conflict and action conflict, respectively. The simulation period was 0-1000 s, and the attack started at 200 s. A new simulation was performed for each

attack.

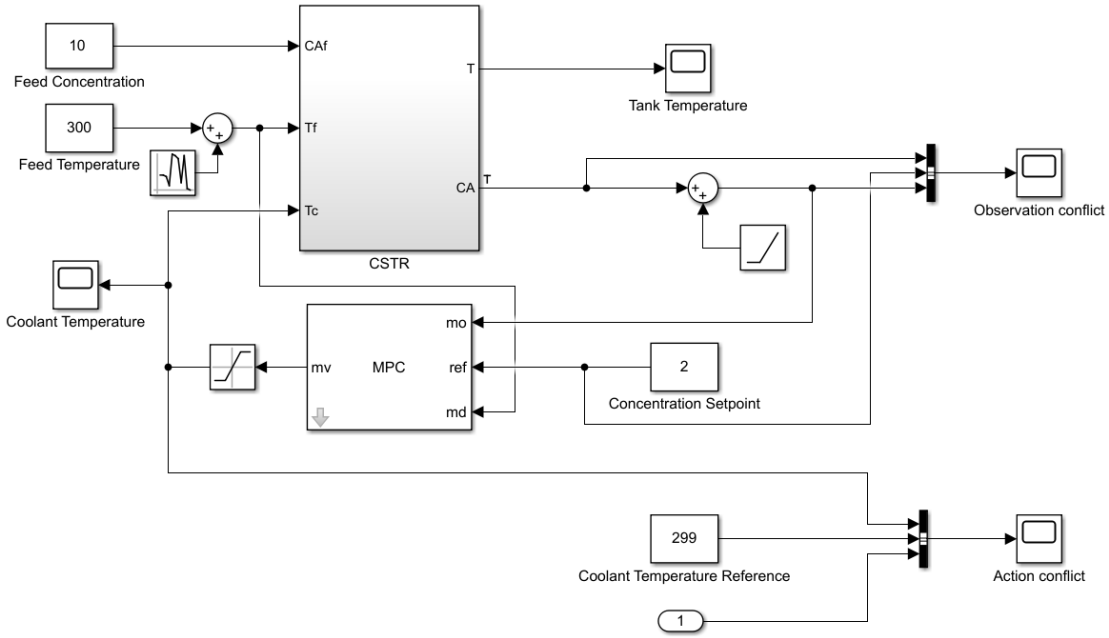

Figure 4.9: Simulation of CSTR under attack (example: FDI on sensor).

- **FDI on sensor**

At 200 s, a cyberattack of FDI on sensor started, simulated by a ramp signal with a slope of 0.002.

$$\tilde{x}(t) = x(t) + 0.002t, t \in [200, \infty) \quad (4.14)$$

- **Setpoint Modification**

Reset the simulation. At 200 s, an attack of setpoint modification started. It modified the concentration from 2 to $2.5 kmol/m^3$.

$$\tilde{x}(t) = x(t) + \ 0.5, \ t \in [200, \infty) \quad (4.15)$$

- **FDI on actuator**

Reset the simulation. At 200 s, an attack of FDI on actuator started, simulated by a ramp signal with a slope of 0.1.

$$\tilde{u}(t) = u(t) + 0.1t, \ \ t \in [200, \infty) \tag{4.16}$$

- **DoS**

Reset the simulation. At 200 s, an attack of DoS started.

$$\tilde{u}(t) = 0\ , \ t \in [200, \infty) \tag{4.17}$$

- **Time delay**

Reset the simulation. At 200 s, an attack of time delay started, and the delay was 100 s.

$$\tilde{u}(t) = \begin{cases} 0, \ \ t \in [200,300), [400,500), \dots \\ u(t), \ \ t \in [300,400), \ [500,600), \dots \end{cases} \tag{4.18}$$

4.3.3. Identify conflicts under attack

Based on the attack representations, the conflict formulations are shown below.

- **FDI on sensor**

$$VOD = 0.002(t - 200), t \in [200, \infty) \tag{4.19}$$

- **Setpoint Modification**

$$VOD = 0.5, \ t \in [200, \infty) \tag{4.20}$$

- **FDI on actuator**

$$VAD = 0.1(t - 200), \ \ t \in [200, 660] \tag{4.21}$$

- **DoS**

$$VAD = -u(t), \ t \in [200, \infty) \tag{4.22}$$

- **Time delay**

$$VAD = \begin{cases} -u(t), \ \ t \in [200,300), [400,500), \dots \\ 0, \ \ t \in [300,400), \ [500,600), \dots \end{cases} \tag{4.23}$$

Since the input and output with MPC would not show a linear relationship, the VOD/VAD under attack cannot be presented directly. The above expressions are the

direct achievable ones, and others could be read in the simulation results.

4.3.3.1. FDI on sensor

For the concentration of the product in the tank (Figure 4.10):

In 0-200 s, both sensor observation and human observation were normal. At 200 s, a cyberattack of FDI on sensor started. In 200-520 s, the operator noticed the observation conflict. The operator considered it a sensor fault and tried to resolve it. At the same time, the operator may not act on the temperature control valve. Since it was not a fault, the fault resolution failed, and then if the operator acted on the valve, the operator tried to decrease the opening of the temperature control valve; however, the MPC controller increased it. An action conflict occurred.

At 520 s, the coolant temperature reached its upper constraint. In the 520 s-1000 s, the observation conflict continued but differed from the former. The operator continued to decrease the opening of the temperature control valve, and the action conflict still existed.

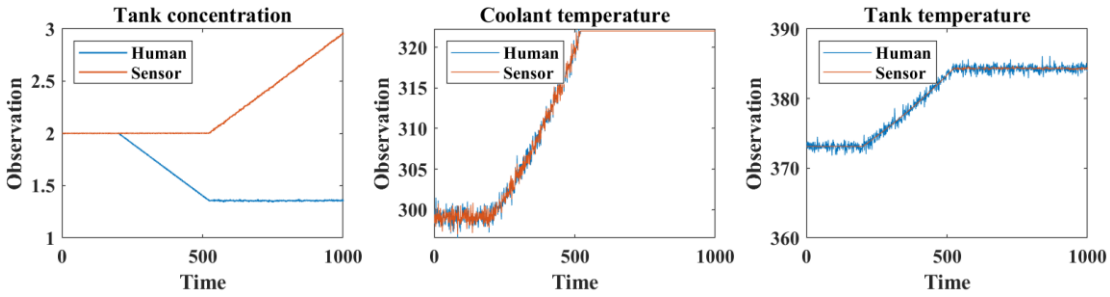

Figure 4.10: Results of FDI on sensor.

4.3.3.2. Setpoint modification

At 200 s, an attack of setpoint modification started. It modified the concentration from

2 to 2.5 $kmol/m^3$ (Figure 4.11). The MPC controller quickly decreased the coolant temperature to 290.2 K. Then the tank temperature decreased to 366.5 K accordingly. The concentration stabilized at 2.5 $kmol/m^3$.

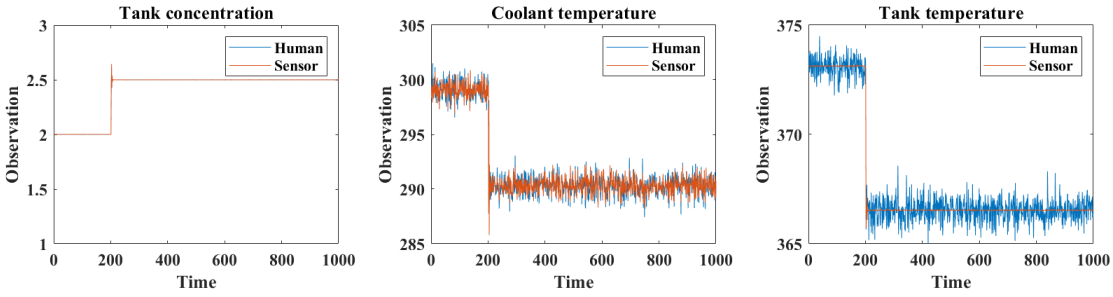

Figure 4.11: Results of setpoint modification attack.

At 200-1000 s, the operator noticed the abnormal situation, and an observation conflict occurred due to the unexpected value. The operator tried to remodify the setpoint. However, the number was locked by the hacker. Then the operator turned to increase the coolant temperature by the valve, but the MPC reverted it rapidly. An action conflict occurred.

4.3.3.3. FDI on actuator

At 200 s, an attack of FDI on actuator started. From the concentration (Figure 4.12), it was normal in 200-430 s, and then started to decrease and stabilized around 1.36 $kmol/m^3$ at 660 s. The human observation and sensor observation on concentration was kept the same.

From the observations of coolant temperature, the human observation started to decrease at 200 s, until the lower limit 276 K at 430 s. The sensor observation started to increase at 430 s until the upper limit 322 K at 660 s.

From the tank temperature, it started to increase at 430 s and stabilized around 384.2 K

at 660 s. In 200-430 s, the operator noticed the abnormal situation on the coolant temperature and tried to increase the temperature while the MPC kept normal control. An observation conflict and an action conflict occurred, then continued differently.

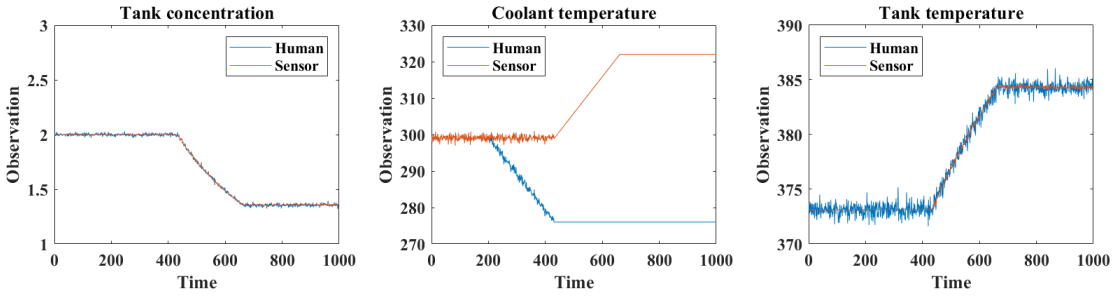

Figure 4.12: Results of FDI on actuator.

4.3.3.4. DoS

At 200 s, an attack of DoS started. The concentration started to fluctuate according to the disturbance (Figure 4.13). The MPC controller lost control capability and kept the reference input. The tank temperature fluctuated due to the disturbance.

In 200-1000 s, the concentration showed unexpected value; therefore, it was an observation conflict between sensor observation and human expectation. The operator took action on the valve; however, the MPC cannot give the expected output. It was also an action conflict.

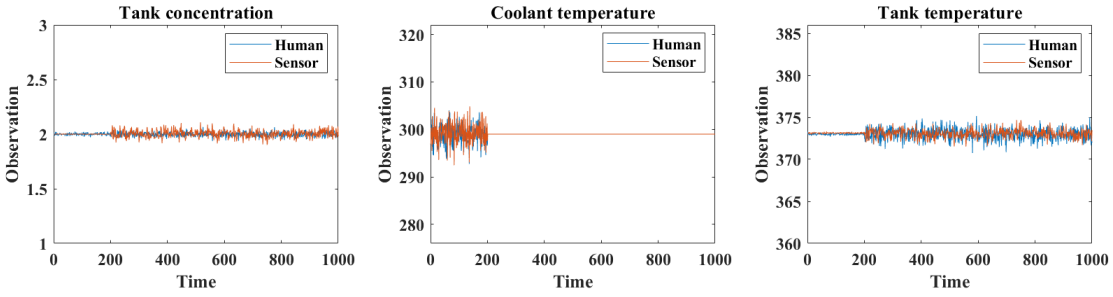

Figure 4.13: Results of DoS.

4.3.3.5. Time delay

At 200 s, an attack of time delay started. The concentration started to fluctuate according to the disturbance periodically every 100 s (Figure 4.14). So did the coolant temperature and tank temperature. In the 200-1000 s, it was also an observation and action conflict similar to DoS.

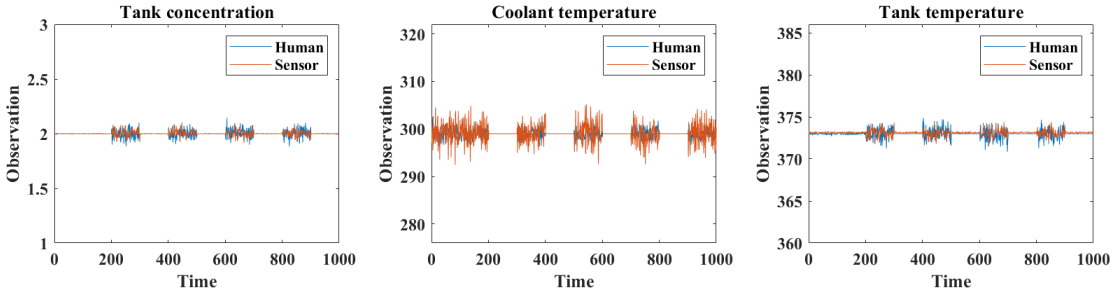

Figure 4.14: Results of time delay attack.

4.3.4. Explain conflict with game paradigm

For the FDI on sensor, it can be inferred that at 520 s, it was the Pareto optimal point (Figure 4.10). Before that, the MPC could adjust the input variable (coolant temperature) in the allowable range to resist the disturbance by the cyberattack, and it did not require manual control by the human operator. 200-520 s was the critical period to judge whether it was a fault or a cyberattack, then take the corresponding resolution.

Similarly, for FDI on actuator, 430 s was the Pareto optimal point (Figure 4.12). From the simulation results, MPC had a buffer effect on FDI attacks, while it cannot withstand setpoint modification, DOS, and time delay.

4.3.5. Assess conflict probability, severity, and risk

This study took the FDI on actuator as an example and calculated the probability, severity, and risk at 200-600 s with Equation (4.11), Equation (4.12), and Equation

(4.13). The probability and risk were shown in Figure 4.15. Five time steps were selected for analysis (Table 4.5).

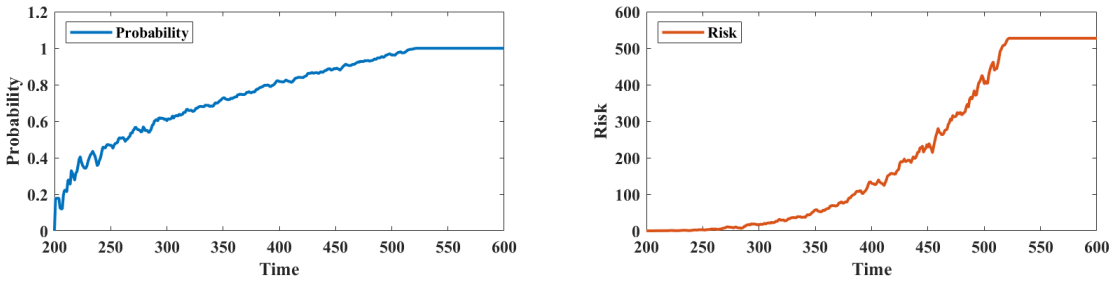

Figure 4.15: Probability and risk results.

Table 4.5: The probability, severity, and risk at sampling time steps.

Variable	257s	300 s	400 s	458s	500 s
Probability	0.50	0.62	0.84	0.92	0.97
Severity	6.6	30.7	181.1	287.8	438.6
Risk	3.3	19.1	151.4	264.5	425.0

Based on the conflict probability, at 257s, it reached 0.50. After that, the conditional probability of an attack was more significant than the conditional probability of a fault; therefore, attack resolution should be considered after 257s.

Based on the risk of conflict, at 458s, it reached half of the maximum risk. This was also a breakpoint to switch to attack resolution.

In this CSTR, compared with the whole range and sampling time steps, 257s would be a relatively aggressive time step to disconnect the network to defense in case it was a cyberattack. At 458s, it was more reasonable to take action for attack resolution.

4.4. Discussion

This study extends human-automation conflict under cyberattacks beyond the former circumstance of sensor faults. Some issues are worth discussing.

Why is it necessary to transfer IT language to process language? The key reason and difficulty are that human operators may not have adequate knowledge about IT technology, especially cyberattack tactics and techniques. This study depicts five generalized cyberattack methods on a closed-loop control system. It is a new attempt to interpret cyber security issues from the perspective of process engineering and control engineering, facilitating safety professionals, engineers, and operators to understand the sophisticated IT language better.

Is the ultimate goal of a cyberattack to change the system input? From the above derivation and analysis (Section 4.2.2), it might be true that the goal is the wrong action to maximize the impairment. This study links the cyberattack expression with conflict expression mathematically. Further, it proves that the attack vector is just the action conflict. It means once the attack succeeds, an action conflict is unavoidable. Hence, it brings new considerations and solutions to the cyber security risks and associated conflicts.

How about the resistance effect to cyberattacks of advanced control? From the application in the CSTR, the MPC has a specific resistance to FDI attacks, while little response to setpoint modification, DoS, and time delay attacks. This might reveal that advanced control can buffer limited and bounded disturbances at the beginning of the cyberattack. Still, it cannot generate enough resistance when unpredictable and unbounded intentional disturbances emerge due to continuous intensified attacks. In addition, the attack on the physical system could be deferred by MPC; however, the

attack on the network system would show an immediate impact, for example, DoS and time delay. Hence, further studies would be deserved on whether other advanced or AI-based control could better resist the cyberattack.

How to distinguish a fault and an attack? Cyberattacks have similar phenomena with faults, in the meantime, fault resolution and cyberattack resolution are independent, and the traditional fault resolution often has no valid response to attacks. That is why the conflict risk assessment model is proposed to switch the strategy. Nevertheless, some engineering concepts are consistent, such as data-driven prediction, physical isolation, and abnormal tolerance. These would incorporate the resolution between faults and cyberattacks.

How to win the game for the operator? Since conflicts under attack are more likely a game between the operator and hacker, this study utilizes some assumptions to reach a solution in this game. The proposed cooperative Pareto paradigm assumes that the hacker and the operator have the same goals to reach an equilibrium. Nevertheless, the operator's best strategy is to avoid this breakeven point.

4.5. Conclusions

This study continues the exploration of human-automation conflict under attack, besides sensor fault, and applies game theory to illustrate conflict. As a result, it presents the mathematical expressions of five common cyberattacks, demonstrating that cyberattacks could cause conflicts and attack strength directly determines the strength of action conflict. In addition, the proposed cooperative Pareto paradigm and the

simulation reveal that control actions could buffer some impact of attacks within a limited range, yet only for FDI attacks. Therefore, conflict risk assessment is required for risk decisions, such as distinguishing a fault and an attack and switching the resolution strategy.

This study also has limitations and further improvement directions. First, cyberattacks are often a combination of multiple tactics in a sequential procedure, and the proposed mathematical expressions are the single ideal illustration and verified in a mature case (CSTR). Further research on mathematical models and simulations of integrated attacks on complex systems is essential. In addition, this study applies Pareto optimal paradigm, yet other paradigms might be more persuasive for such a dynamic and uncooperative game. This starts with incorporating game theory and conflict theory to confront cyberattacks. More game-theoretic paradigms can be examined and applied in the future. Lastly, this study proposes process solutions to the conflict rather than IT solutions. Undoubtedly, the combination of IT techniques should also be considered, such as intrusion detection and vulnerability scanning. Therefore, the resolution of human-automation conflict relies on the joint effort of process engineers and IT engineers. It is still a long way to explore cyber threats from process safety perspectives comprehensively.

Chapter 5: Conflict from Situation Awareness

Preface

A version of this chapter has been published in *Industrial & Engineering Chemistry Research*. I am the primary author, along with the co-authors, Drs. Faisal Khan, Md. Tanjin Amin, Salim Ahmed, Syed Imtiaz, and Stratos Pistikopoulos. I developed the conceptual framework for the methodology and carried out the literature review and case study. I prepared the first draft of the manuscript and subsequently revised the manuscript based on the co-authors' and peer review feedback. Co-author Dr. Faisal Khan helped develop the concept, verify the methodology, review, and revise the manuscript. Co- authors Drs. Md. Tanjin Amin, Salim Ahmed, Syed Imtiaz, and Stratos Pistikopoulos provided support in implementing the concept and verifying the methodology. The co-authors provided fundamental assistance in validating, reviewing, and correcting the methodology, case study, and results. The co-authors also contributed to the review and revision of the manuscript.

Reference: Wen, H., Amin, M. T., Khan, F., Ahmed, S., Imtiaz, S., & Pistikopoulos, E. (2023). Assessment of Situation Awareness Conflict Risk between Human and AI in Process System Operation. *Industrial & Engineering Chemistry Research*, 62(9), 4028–4038. https://doi.org/10.1021/acs.iecr.2c04310

Abstract

The conflict between human and artificial intelligence is a critical issue, which has recently been introduced in Process System Engineering, capturing the observation and

action conflicts. Interpretation conflict is another source of potential conflict that can cause serious concern for process safety, as it is often perceived as confusion, surprise, or a mistake. It is intangible and associated with situation awareness. However, interpretation conflict has not been studied with the required emphasis. The current work proposes a novel methodology to quantify interpretation conflict probability and risk. The methodology is demonstrated, tested, and validated on a two-phase separator. The results show that interpretation conflict is usually hidden, mixed, or covered by traditional faults, and noises in observation and interpretation, including sensor faults, logic errors, cyberattacks, human mistakes, and misunderstandings, may easily trigger interpretation conflict. The proposed methodology will serve as a mechanism to develop strategies to manage interpretation conflict.

Keywords: Human-AI conflict, interpretation conflict, noise, automation, digitalization.

5.1. Introduction

The inception of Industry 4.0 and the enhanced use of digital technologies and digitalization are reshaping the operation and structure of process systems (Bequette, 2019; Vaccari et al., 2021). Effective utilization of artificial intelligence (AI) and machine learning (ML) algorithms is pivotal to ensure the successful adoption of Industry 4.0 and digitalization, and day by day, these technologies are increasingly being used in process industries (Udugama et al., 2020). These have significantly improved the performance of sensors and controllers, model prediction accuracy, parameter estimation, and process optimization (J. H. Lee et al., 2018). Although

process performance has notably improved due to the use of AI and ML, process safety incidents are still occurring (J. Lee et al., 2019). Excessive dependence on AI-based automated technologies may result in accidents (e.g., the 2005 Buncefield fire in the UK) (Khan et al., 2021). However, research on industrial automation and AI is still a key area intending to assist humans in decision-making (Ghosh & Wayne Bequette, 2020; Sokolov, 2020).

Despite a growing focus on industrial automatization, human beings are yet in the loop for ensuring safety, especially in petrochemical industries (Wanasinghe et al., 2021). The role of AI, in most cases, is to help operators to have a better prediction of the situation. For instance, operators rely on ML algorithms to narrow down the search window in root cause diagnosis to restore the process to normal operating mode due to an abnormal operating condition (Lu et al., 2019; Patwardhan et al., 2019). Therefore, deeply studying the interaction and collaboration in process systems is necessary. More specifically, it is of utmost importance to understand how AI decides and predicts the present and future by judging its surroundings and using the in-built logic.

Situation awareness (SA) – an appropriate awareness of the situation – plays a crucial role in the performance of humans and AI agents (Smith & Hancock, 1995). For instance, SA concept is widely used for aircraft safety in the aviation industry, and ill-judging a situation is one of the major reasons for aviation accidents (Stanton et al., 2001). Although the definition of SA varies from the different scholars' perspectives (Stanton et al., 2017), SA is often categorized in the domain of human factor and

considered a performance-related psychological concept. A widely accepted proposition is Endsley's three-level model: perception (level 1), comprehension (level 2), and projection (level 3) (Endsley, 1995). Further studies also emphasize descriptions of individual performance at an abstract and general level (Endsley, 1988a), with rare quantification (Endsley, 1988b). The extensive research is team SA models discussing information exchange and team cooperation (Salas et al., 2017). However, the focus of research has always been on humans, and the SA of automated systems has rarely been studied. Recent research starts to consider distributed SA (Salmon et al., 2017), which means the entire system-level comprehension or compatible awareness in the human-intelligent distributed system. But none of these discusses the SA differences between humans and AI agents.

In Process System Engineering (PSE), the issue of SA in major chemical accidents has also received attention, and multiple case studies have been conducted (Naderpour et al., 2015). The chemical industries emphasize team SA among operators and engineers since its role in preventing catastrophic accidents is paramount (Kaber & Endsley, 1998). It also stresses the holistic SA and distributed SA at the system level (Nazir et al., 2014). There has also been progress in quantifying the impact of SA, for example, the combination of SA and the Bayesian network (Naderpour et al., 2014). Digital technology has also expanded the research in the field of SA, and scholars have proposed to use eye tracking technology to evaluate SA (Bhavsar et al., 2017). However, no notable study exists on the SA difference quantification between humans and AI in

the chemical industries.

Generally, humans are better in the context of SA due to their ability to recognize new situations earlier than AI. For instance, the recently discovered adversarial attacks showed that adding a small noise could make the AI misclassify the image (Szegedy et al., 2014). Numeric process data can also be contaminated by adding a small noise, which may mislead the logic solver or ML algorithms. Similar attacks would be easier in the form of false data injection (FDI) or denial of service (DoS). The current authors believe this performance degradation is due to AI's loss of SA. It is alarming from a safety perspective since a controversial but widely discussed prediction is that AI will surpass all humans' intelligence by 2045 when technology singularity will occur (Figure 5.1) (Grossman, 2011; Kurzweil, 2005). The studies suggest AI is growing exponentially, while human intelligence is growing slowly and approximately linearly. Due to this exponential growth in AI's intelligence, it is expected to exercise more dependence on AI-based automated systems in process industries. It may be a boon because of their possible improvement in decision-making. However, it also paves the way for experiencing catastrophic accidents due to overreliance on technologies that are poor in the context of SA.

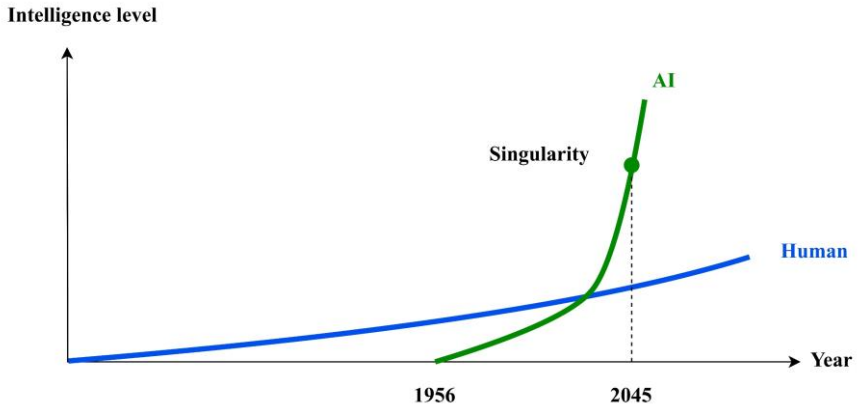

Figure 5.1: Intelligence growth of human and AI.

Interpretation of a situation is an important part of SA. It leads to the decision and actions suggested by the automated systems. For instance, sensors gather information about their surroundings (or situation) and make the in-built logic to interpret the situation and subsequently, take decisions and actions. Similarly, human beings have a mechanism to interpret a situation and make a decision. These two decision-makers may or may not coincide in terms of their interpretation which may result in an interpretation conflict. It is worth noting that conflict analysis in process systems between humans and machines is a novel concept that has recently been proposed (Wen et al., 2022). However, the authors have shown how to identify and assess risk due to observation and action conflicts without going into a detailed analysis of the interpretation conflict. Associated with SA, interpretation conflict is also from the cognition perspective and is usually intangible. It is more likely to occur in cases of logic errors, human misunderstandings, and cyberattacks. Even when the object is imperfect or mixed with noises, an observation conflict may occur that can trigger an interpretation conflict. In the aviation industry, situations of interpretation conflict are mode confusion

(Hamburger, 1966; Rushby, 2002) and automation surprise (Combéfis et al., 2016; Woods & Sarter, 1998). Also, in the context of self-driving cars, unexpected braking or changing lane confuses the driver (Moscoso Paredes et al., 2021), and the driver may have situational reactions under stress; this is an example of interpretation conflict.

Currently, the interpretation conflict between humans and AI has not been properly addressed. Though mode confusion and automation surprise have some compelling works, these studies start from the macro level, focusing on the action, not the recognition process. To the authors' best knowledge, no work on PSE has focused on demystifying interpretation conflict between humans and AI from a safety perspective (e.g., shutting down the operation to avoid acknowledging the significant risk due to an interpretation conflict). To eliminate this gap, this study attempts to answer the following research questions: (i) what is interpretation conflict? (ii) how to identify interpretation conflict? (iii) how does interpretation conflict occur? (iv) how to assess the risk due to an interpretation conflict?

In addition, this study proposes a methodology to reveal the interpretation conflict and applies it in a two-phase separator. The novelties of this paper are the following: (i) introducing the concept of interpretation conflict, (ii) deconstructing the evolution process of interpretation conflict, (iii) exploring the impact of various noises on interpretation conflict, and (iv) developing a novel methodology to assess the risk as a result of an interpretation conflict.

The paper is arranged as follows: Section 5.2 describes the interpretation conflict and

its evolution. Section 5.3 presents the proposed novel methodology to assess interpretation conflict risk. The simulation and application are described in Section 5.4. The results are discussed in Section 5.5. Finally, conclusions and future directions are shown in Section 5.6.

5.2. Situation awareness conflict (interpretation conflict) evolution

5.2.1. Definition

Conflict has been defined as *the difference in the observation, interpretation, or action of one or more variables by different participants* (Wen et al., 2022). Therefore, interpretation conflict is the difference in interpretation by different participants. Figure 5.2 shows the recognition process of AI to imitate human recognition.

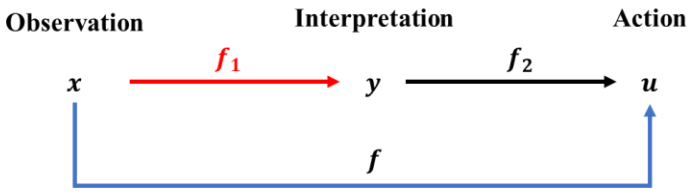

Figure 5.2: Recognition process of AI.

Where x is observation, y is interpretation, u is action, f is the function from observation to action, f_1 is the function from observation to interpretation, and f_2 is the function from interpretation to action.

The traditional control theory solves f with the state space equation. f can be destructed into two subfunctions, f_1 and f_2. Action is usually one-on-one with interpretation results; hence, in this study, the research focus is f_1, which is situation awareness. Therefore, the fundamental cause of interpretation conflict is the difference

of f_1 between humans and AI.

As mentioned earlier, humans have a better understanding of newer situations compared to AI. Except for confirmed correct and convinced wrong, there are some grey areas of human feeling, which is the deviation between human interpretation and AI interpretation; for example, mode confusion or automation surprise, and similar feelings include hesitation, doubt, and unsureness. The relationship between such deviations and interpretation conflict is shown in Figure 5.3. When it is confirmed that AI is making an accurate interpretation, there should be no interpretation conflict. Otherwise, any confusion, surprise, and convinced wrong can be categorized as interpretation conflict. It is worthwhile to mention that current work is assessing interpretation conflict from a human perspective since humans have a better SA at the current level of intelligence.

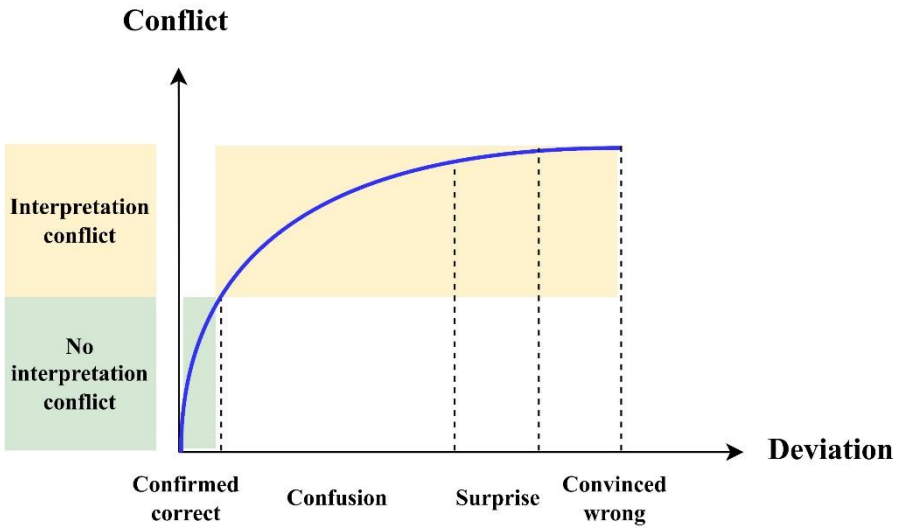

Figure 5.3: Relationship between human feeling and interpretation conflict.

5.2.2. Evolution process and mathematical formulation

Although the emergence of interpretation conflict may be instant, it still has a deconstructable evolution process. First, the conflict variables are defined: (i) variable

of observation difference (VOD) is the difference in observation of process value from different observers; (ii) variable of interpretation difference (VID) is the difference in interpretation of process value from different interpreters; (iii) variable of action difference (VAD) is the difference in control action by different participants (Wen et al., 2022).

In a perfect situation without noise, there is no observation conflict, no interpretation conflict, and no action conflict, which means $VOD = 0$, $VID = 0$, and $VAD = 0$. As different noises work on AI and humans, in most cases, there should be differences, and these are the basic causes of a conflict. Figure 5.4 describes how human-AI interpretation conflict occurs.

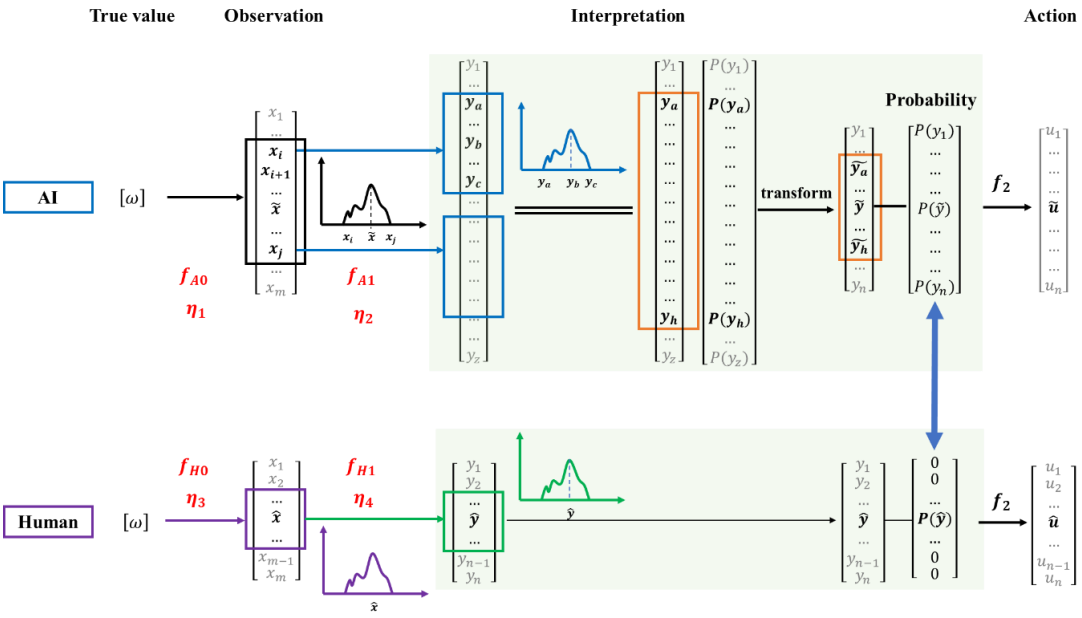

Figure 5.4: Interpretation conflict between AI and human.

Where the subscript A stands for AI, the subscript H stands for human; ω is a supposed true value; f_0 is the function from true value to observation, f_{A0} is the function from true value to sensor observation, f_{H0} is the function from true value to human

observation; f_{A1} is the function from observation to interpretation of AI, f_{H1} is the function from observation to interpretation of human; η denotes noise, η_1 is the noise in sensor observation, η_2 is the noise in AI interpretation, η_3 is the noise in human observation, η_4 is the noise in human interpretation; m is the observation vector size, n is the action vector size and the transformed interpretation vector size, z is the full size of the extended interpretation vector, the other lowercase letters from a to z represent the subscripts of interpretation or observation; $\tilde{x}$ is the most possible sensor observation, $\tilde{y}$ is the most possible AI interpretation, $\tilde{u}$ is the most possible AI action; $\hat{x}$ is the most possible human observation, $\hat{y}$ is the most possible human interpretation, $\hat{u}$ is the most possible human action.

For AI, given a true value, ω, as there are noises, η_1 in observation to affect f_{A0}, and the noises can be measurement error by sensor, sensor fault, or FDI on sensor; also, there are noises, η_2 in interpretation, which affects f_{A1}, and the noises can be logic error, adversarial attack, FDI on controller, or DoS. Therefore, the observation, interpretation, and action equations of AI will be

$$x = f_{A0}(\omega) = \tilde{x} \in [x_i, x_{i+1}, \dots, x_j] \tag{5.1}$$

$$y = f_{A1}(x) = \tilde{y} \in \left[[y_a, \dots, y_b, \dots, y_c], [y_d, \dots, y_e], \dots [y_g, \dots, y_h]\right] = [y_a, \dots, y_h] \tag{5.2}$$

$$u = f_2(\tilde{y}) = \tilde{u} \tag{5.3}$$

As one observation may correspond to multiple possible interpretations, the interpretation vector becomes longer; then it needs to be transformed to the same size as the action vector. The range $[y_a, \dots, y_h]$ will be transferred to $[\widetilde{y_a}, \dots, \widetilde{y_h}]$.

For humans, usually, SA is an instant and straightforward process. Humans may have an estimated range of observations and then give the most possible guess; similarly, humans will give several corresponding interpretations, and then make a clear choice directly, though the whole process is unknown. As there is noise, η_3 in observation, and the noise is mostly human mistake or measurement error (by equipment or eyes); there is noise, η_4 in interpretation, and the noise is mostly human misunderstanding. Therefore, the observation, interpretation, and action equations of human will be

$$x = f_{H0}(\omega) = \hat{x} \tag{5.4}$$

$$y = f_{H1}(\hat{x}) = \hat{y} \tag{5.5}$$

$$u = f_2(\hat{y}) = \hat{u} \tag{5.6}$$

Consequently, VOD, VID, and VAD can be represented by Equation (5.7), Equation (5.8), and Equation (5.9), respectively.

$$VOD = \tilde{x} - \hat{x} \tag{5.7}$$

$$VID = \tilde{y} - \hat{y} \tag{5.8}$$

$$VAD = \tilde{u} - \hat{u} \tag{5.9}$$

5.3. The proposed methodology to assess interpretation conflict risk

5.3.1. General description

The methodology is shown in Figure 5.5 and detailed steps are described below.

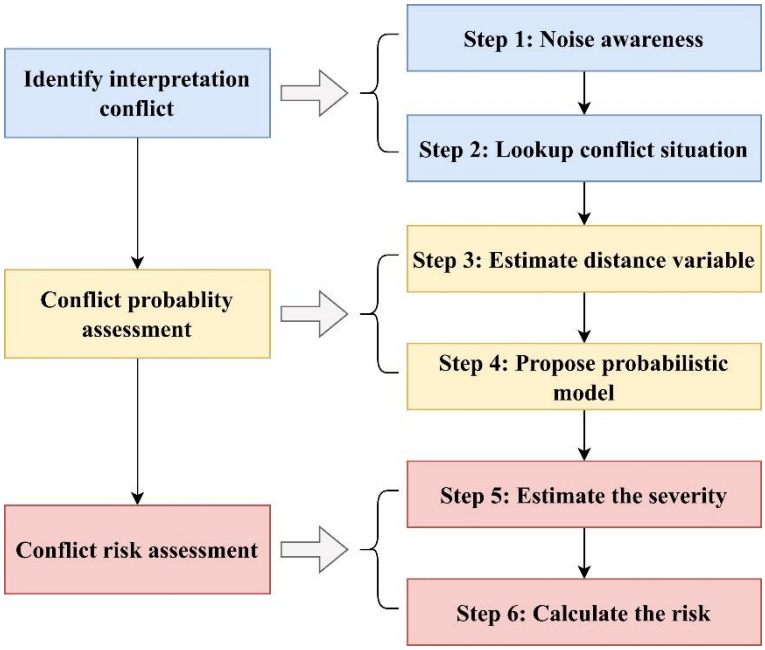

Figure 5.5: Methodology to assess interpretation conflict risk.

Step 1: To identify interpretation conflict, first, it is necessary to monitor the process value and be aware of noises, including sensor faults, logic errors, measurement errors, cyberattacks, mistakes, and misunderstandings.

Step 2: In this step, the situations of interpretation conflict are categorized and summarized, and then the lookup method is applied to identify the conflict situations.

Step 3: Based on Bayesian theory and fitted triangular distribution, the interpretation probability is derived. The distance between the vector of AI interpretation probability and the vector of human interpretation probability is measured.

Step 4: The probabilistic model of interpretation conflict is developed in this step.

Step 5: After analyzing the severity distribution, the equation of conflict severity is proposed.

Step 6: The risk is quantified and graded for decision-making.

5.3.2. Identify interpretation conflict

5.3.2.1. Noise awareness

The interpretation conflict can occur in an instant and is coupled with observation conflict. Usually, the noises could be reflected in the abnormal process values. Hence, the operator is required to monitor any fluctuations and deviations, and be aware of sensor faults, logic errors, cyberattacks, measurement errors, mistakes, and misunderstandings. In this study, noise is a broader collective term, which may include white Gaussian noise, random noise, perturbation, disturbance, interference, and error.

5.3.2.2. Lookup conflict situation

The lookup method is applied to identify interpretation conflict situations. The classification of conflict situations is shown in Figure 5.6. In the perfect situation, there is no noise in observation and interpretation, therefore, it is a normal operation without conflicts (Situation 1). Interpretation conflict may arise from noise in interpretation (e.g., logic error or human misunderstanding) (Situation 2). If there is noise in observation, such as measurement error, sensor fault, or human mistake, there may be observation conflict; consequently, it triggers interpretation conflict (Situation 3). When it is small enough ($VOD < \pm\sigma$), it is acceptable. In some cases, observation noise and interpretation noise may exist together, and the interpretation conflicts overlap (Situation 4). In summary, Situations 2, 3, and 4 are interpretation conflicts.

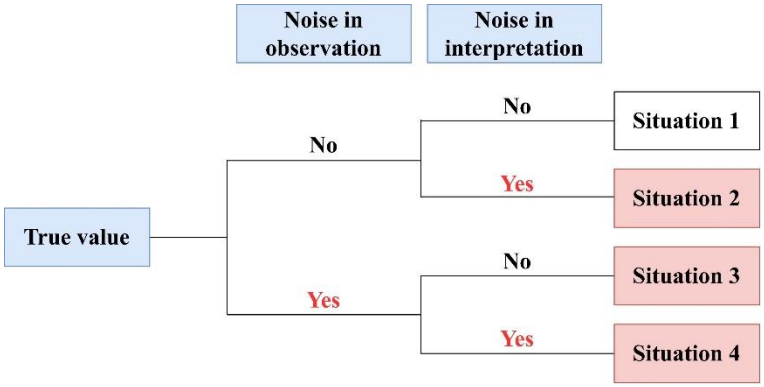

Figure 5.6: Situations of interpretation conflict.

5.3.3. Conflict probability assessment

5.3.3.1. Estimate the distance variable

Suppose the observations have a range that a triangular distribution can fit (Figure 5.7).

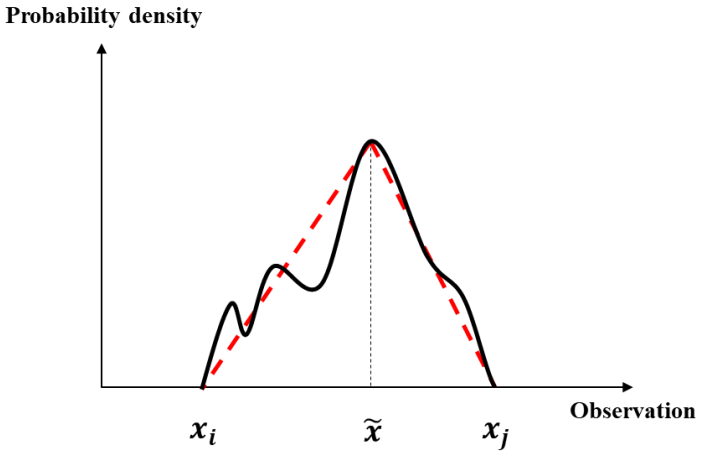

Figure 5.7: Fitted triangular distribution of observations.

Then the probability of each observation is

$$P(x_i) = \frac{e^{PDF(x_i)}}{\sum_0^m e^{PDF(x_i)}} \tag{5.10}$$

Where PDF is the probability density function of observations.

Similarly, the probability of each interpretation can be estimated as

$$P(y_k|x_i) = \frac{e^{PDF'(y_k)}}{\sum_0^z e^{PDF'(y_k)}} \tag{5.11}$$

Where PDF' is the probability density function of interpretations, which is another triangular distribution.

The observation determines what the interpretation will be. Therefore, the interpretation result follows the conditional probability rule, and the interpretation probability is

$$P(y \cap x) = P(x)P(y|x) \tag{5.12}$$

For ease of understanding, $P(y \cap x)$ is simplified as $P(y)$; for example, $P(y_k) = P(x_i)P(y_k|x_i)$. After transforming to the same size as the action vector, the final vector of AI interpretation probability can be obtained. On the other hand, for humans, it has $P(\hat{y}) = P(\hat{x})P(\hat{y}|\hat{x})$. The probabilities of other interpretation results are marked as 0 to form the vector of human interpretation probability.

Here it is proposed to measure the distance d between the vector of AI interpretation probability and the vector of human interpretation probability.

$$VID \propto d = cross\ entropy(P(y_A), P(y_H)) \tag{5.13}$$

Also, VID varies to d. The cross-entropy is widely applied in deep learning and is more significant compared with other distance algorithms in this study.

As noise is usually time-varying and the interpretation conflict often lasts for a period, the range of observations may vary from one time step to another. Hence, at each time step, the AI observation function and interpretation function should be different. This statement is also valid for human observation and interpretation. Therefore, for multiple observations in time series, the distance varies with time (Figure 5.8).

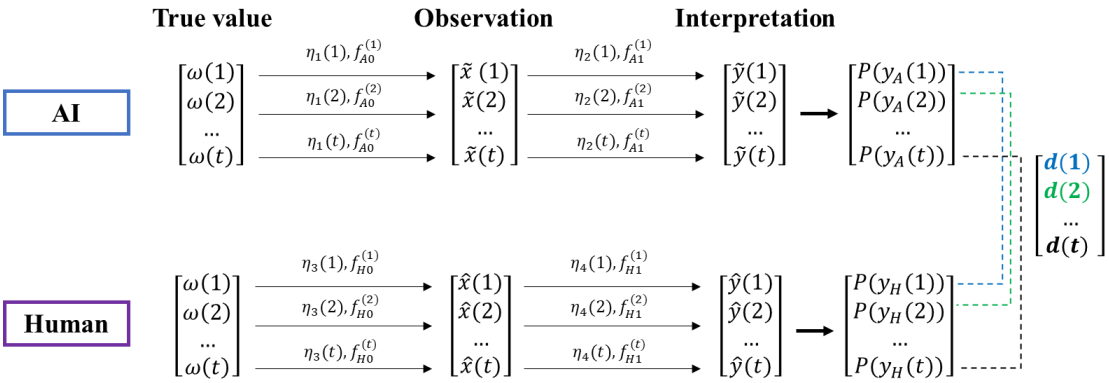

Figure 5.8: Distance variable of interpretation conflict.

5.3.3.2. The proposed probabilistic model

Based on the above derivation, when d=0, there is no interpretation conflict. There should be a maximum d_{max}, when $d = d_{max}$, an interpretation conflict certainly occurs. However, when $0 < d < d_{max}$, there is a possibility to occur an interpretation conflict (Figure 5.9).

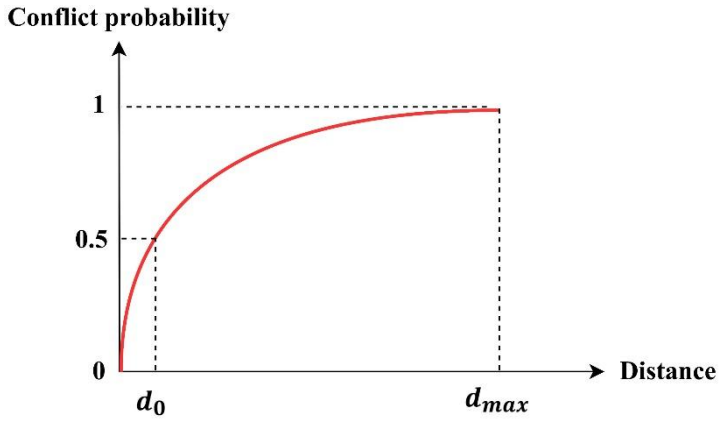

Figure 5.9: Probability distribution of interpretation conflict.

Therefore, the interpretation conflict probability, P is proposed as

$$P = BETA.INV(\frac{d}{d_{max}}, \alpha, \beta) \tag{5.14}$$

Where α and β are the parameters of the beta inverse distribution. d_0 responds $P = 0.5$; d_{max} is associated with the size of the vector, which is the vector distance

when the AI and human give different interpretations with 100% confidence, for example,

$$d_{max,1} = cross\ entropy([1],[0]) = 36.04 \tag{5.15}$$

$$d_{max,2} = cross\ entropy\left(\begin{bmatrix}1\\0\end{bmatrix},\begin{bmatrix}0\\1\end{bmatrix}\right) = 18.02 \tag{5.16}$$

Table 5.1 shows some examples of d_{max}.

Table 5.1: Example of maximum d.

Vector size	**Distance**
1	36.04
2	18.02
3	12.01
…	…
100	0.36
…	…
1000	0.04

5.3.4. Conflict risk assessment

5.3.4.1. Estimate the severity

The conflict severity, S is proposed as

$$S = \begin{cases} \frac{d}{d_0}, 0 < d \leq d_0 \\ e^{\sqrt[n]{d-d_0}}, d_0 < d \leq d_{max} \end{cases} \tag{5.17}$$

When $0 < d \leq d_0$, the severity follows a linear function; and at d_0, the severity is 1; when $d_0 < d \leq d_{max}$, the severity follows an exponential function (Figure 5.10).

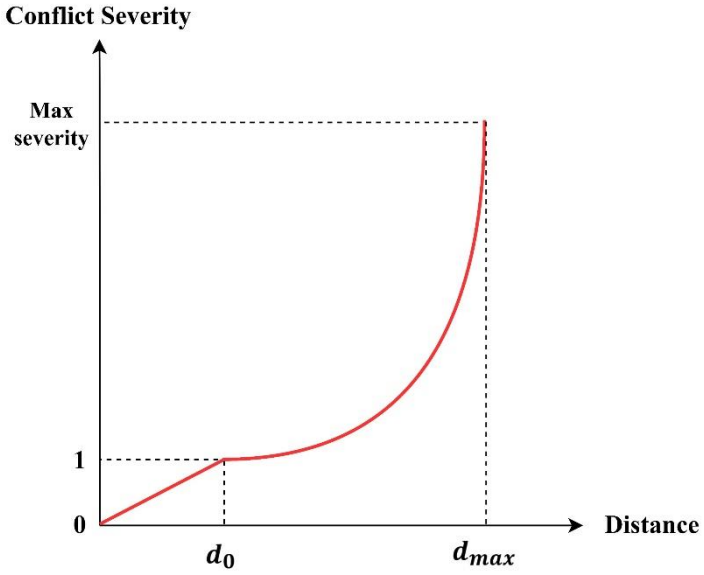

Figure 5.10: Severity distribution of interpretation conflict.

5.3.4.2. Calculate the risk

Correspondingly, the risk, R is

$$R = P \times S \tag{5.18}$$

The risk can be graded in two categories. When the risk is less than 0.5, the interpretation conflict is acceptable. Otherwise, it is alarming, and action needs to be taken to minimize the risk.

5.4. Application of the proposed methodology

5.4.1. Case description and simulation

The two-phase separator is a common unit to separate oil and gas (Figure 5.11). This study sets two types of level measurement: a tubular level gauge and a *differential pressure transmitter. A*n operator monitors the system by reading the tubular level gauge. The differential pressure transmitter is connected to the level controller and the control

valve.

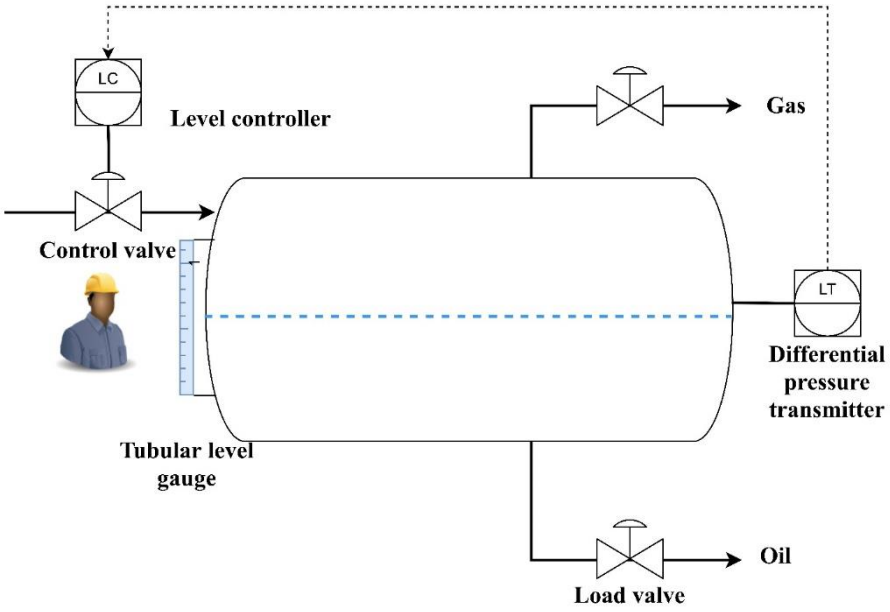

Figure 5.11: Two-phase oil and gas separator.

This study assumes that crude oil has the same density as water. A built-in model in MATLAB is used and the detailed assumptions can be found in the references (The MathWorks, 2022). The cross-sectional area of the tank, setpoint height of oil in the tank, responding valve opening, the height of the tank, cross-sectional area of the pipe, and maximum inflow rate of oil intake are 1 m^2, 0.50 m, 50%, 1 m, 0.005 m^2, and 1 m^3/s, respectively. The variables and ranges are presented in Table 5.2. N denotes normal distribution.

Table 5.2: Variables of the two-phase separator.

Variable	Symbol	Description	Range
Input	u	Action: the valve opening	[0,100%]
State	x	Observation: the height of oil	[0,1]; $x \sim N(0.5, 0.01^2)$
Output	y	Next time step observation	[0,1]; $y \sim N(0.5, 0.01^2)$

For the two-phase separator, the differential equation is

$$\frac{dV}{dt} = C\frac{dh}{dt} = bq - a\sqrt{h} \tag{5.19}$$

Where V is the volume of oil in the tank, C is the cross-sectional area of the tank, h is the height of oil in the tank, b is a constant related to the flow rate into the tank, q is the inlet flow rate, and a is a constant related to the flow rate out of the tank.

For the subsystem of the inlet valve, it has

$$\frac{dq}{dt} = K_u u \tag{5.20}$$

Where K_u is the coefficient constant of the valve opening.

Referring to the built-in model in MATLAB,(The MathWorks, 2022) the transfer function from the input variable to the output variable in our case is proposed as

$$G(s) = \frac{0.8}{2s^2+s} \tag{5.21}$$

The simulation setup in the MATLAB/Simulink R2021a environment is shown in Figure 5.12. A proportional-integral-derivative (PID) controller simulates the AI, and a proportional controller simulates the human. The techniques to simulate noises are random number signals representing measurement errors, input table with manipulated observations serving as the sensor fault, addition and subtraction of constant numbers working as human mistakes, and switch modules with different values representing the logic error and human misunderstanding.

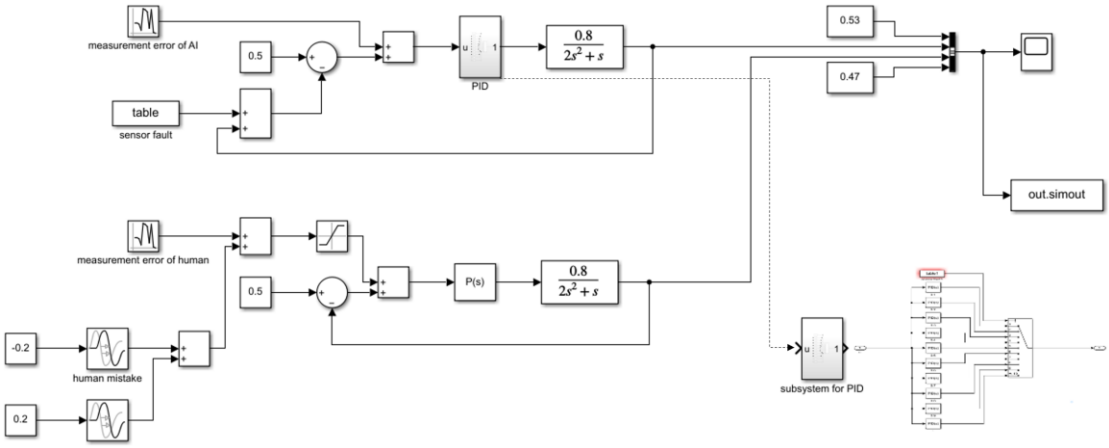

Figure 5.12: Simulink model of interpretation conflict.

5.4.2. Identify interpretation conflict

5.4.2.1. Noise awareness

As mentioned earlier, the initial task is to find the noise in observation and interpretation.

The simulation steps follow the description in Table 5.3 to add the noise gradually.

Table 5.3: Simulation steps to add noises.

Time	Noise type
1-500 s	No noise
501-1000 s	A sensor measurement error with Gaussian white noise $N(0, 0.001^2)$
1001-1500 s	A human measurement error with Gaussian white noise $N(0, 0.01^2)$
1501-2500 s	A sensor fault to manipulate observations with a triangular distribution [0.2, 0.7, 0.9]; also, an observation mistake by human with -0.2 of each observation at 1701-1710 s
2001-3000 s	A logic error on the PID controller to manipulate proportional value with a triangular distribution [0,8,10] (default is 0.2)
2501-3000 s	A misunderstanding on the P controller to change the proportional value to 1 (default is 0.2)

The simulation results are shown in Figure 5.13. The sharp variation in the first 80 s is the initial fluctuation to reach a stable state.

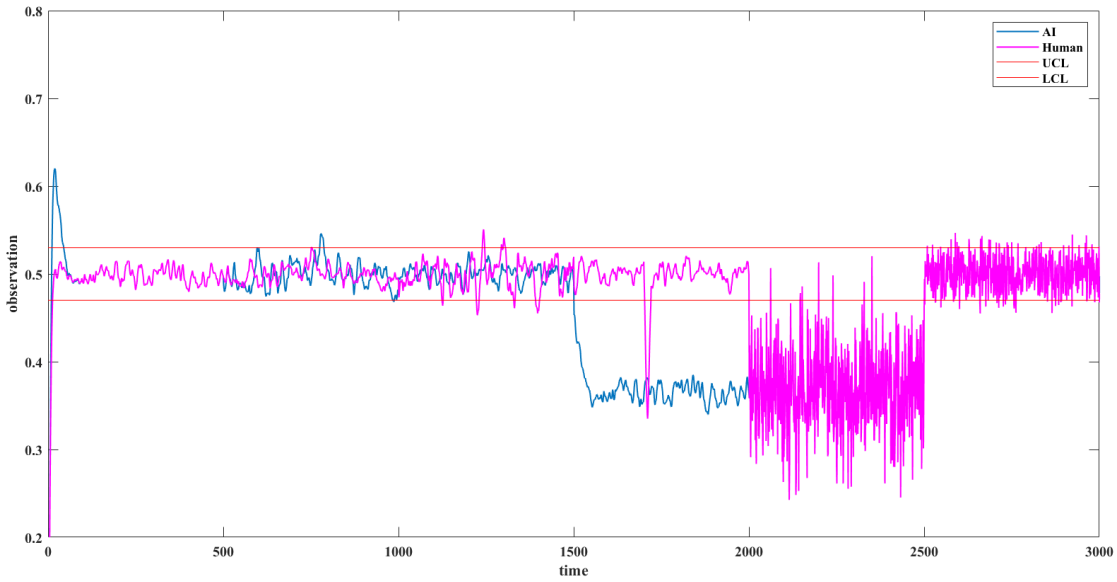

Figure 5.13: The observations of the oil level.

Correspondingly, the VOD for observation conflict is obtained (Figure 5.14).

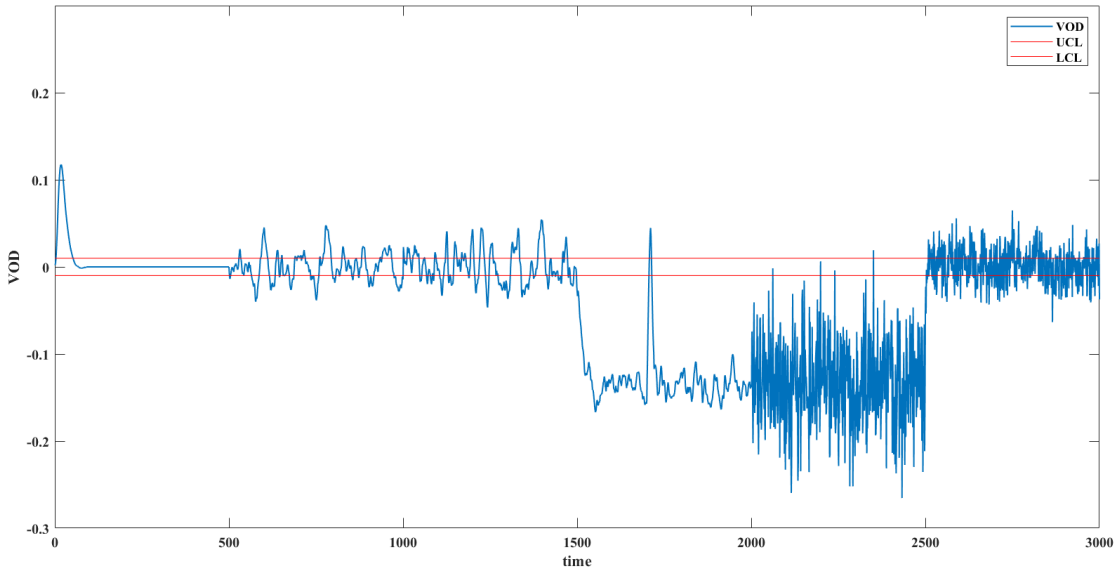

Figure 5.14: VOD for observation conflict.

5.4.2.2. Lookup conflict situation

The situations of interpretation conflict are identified and summarized in Table 5.4.

Table 5.4: Identification results of interpretation conflict.

Time	Conflict situation	Interpretation conflict
1-500 s	Situation 1	No
501-2000 s	Situation 3	Yes
2001-2500 s	Situation 4	Yes
2501-3000 s	Situation 2	Yes

Situation 1: In 0-500 s, as there is no noise at any time, human observation and sensor observation are the same. There is no interpretation conflict.

Situation 3: In 501-1000 s, as there is a sensor measurement error, the sensor observations deviated from the true values. However, they are still mostly between control limits. In 1001-1500 s, the human measurement error makes observations deviate from the reference values. In both periods, observation conflict persists for a few instances that may trigger interpretation conflict. This is a situation, which is described as confusion.

In 1501-2000 s, a sensor fault with a triangular distribution [0.2, 0.7, 0.9] is added. As most observations are higher than the setpoint 0.5, the controller takes action to adjust. It results in a low liquid level. It causes observation and interpretation conflicts. This is an automation surprise.

In 1701-1710 s, an observation error happens from the operator end, which makes the observation curve sharply deviate from the true value. In most cases, such a mistake stays for a short period, and the operator may become aware of it later. Observation and interpretation conflicts also occur in such situations.

Situation 4: In 2001-2500 s, a logic error on the PID controller to manipulate proportional value with a triangular distribution [0,8,10] (default is 0.2) occurs. It makes

the controller lose its accuracy to adjust the liquid level. Together with the sensor fault, observation conflict and overlapped interpretation conflict occur.

Situation 2: In 2501-3000 s, the operator finds the cause of the sensor fault and solves it. However, the fluctuation keeps occurring because the logic error is still present. The operator misunderstands the situation and takes a wrong action, which is simulated by changing the proportional value of the proportional controller to 1 (default is 0.2). Human observations fluctuate beyond the limit. An interpretation conflict occurs.

5.4.3. Conflict probability assessment

5.4.3.1. Calculate the distance in 2001 s

As the logic error occurs in 2001-3000 s, select the time step 2001 s as the research object, where sensor observation is 0.36 and the proportional value of PID is 6. In this simulation, as the switch modules are used to represent the shift between logic decisions (proportional value of PID), therefore, once determined, the proportional value is certain, respectively with a certain probability, and other probabilities are 0. According to Equation (5.10), Equation (5.11), Equation (5.12), and Equation (5.13), each variable and value is calculated and shown in Table 5.5.

Table 5.5: Calculation results at 2001 s.

Variable	Value
$P(x)$	6.2E-03
$P(y\|x)$	1.0E-02
$P(y)$	6.4E-05
$P(y_A(2001))$	$[0,0,\dots,6.4\mathrm{E}-05,\dots,0]^T$
$P(y_H(2001))$	$[0,0,\dots,1.4\mathrm{E}-04,\dots,0]^T$
d_{max}	3.6E-01
d	2.3E-05

5.4.3.2. Calculate the probability in 2001 s

According to Equation (5.14), suppose α is 20 and β is 1, then the conflict probability at time 2001 is

- $P(2001) = 0.62$

5.4.4. Conflict risk assessment

According to Equation (5.17), and Equation (5.18), for the time step 2001 s, the severity and risk are

- $S(2001) = 2.46$
- $R(2001) = 1.52$

The risk is appreciably greater than 0.5. Thus, interpretation conflict occurred in 2001 s, just at the same time when the logic error happened. In addition, the risk of the whole period is calculated and shown in Figure 5.15.

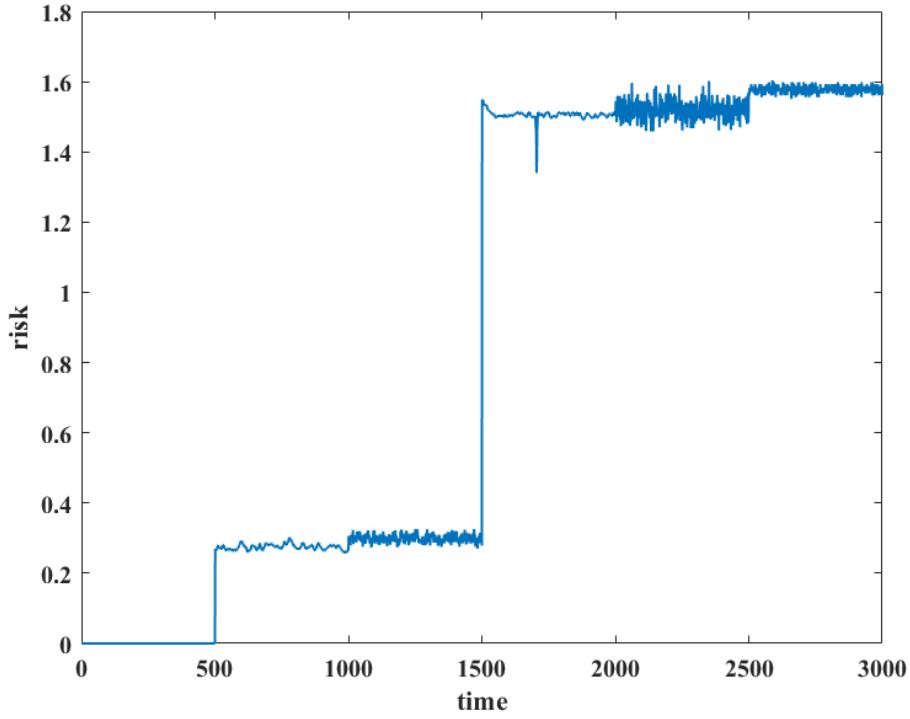

Figure 5.15: Risk in 0-3000 s.

In 501-1000 s (sensor measurement error) and 1001-1500 s (human measurement error), the risk is less than 0.5, which can be considered relatively small, and the risk of interpretation conflict is acceptable. In 1501-2000 s, the sensor fault increases the risk sharply. When it overlaps with logic error in 2001-2500 s, the risk increases even higher. In 2501-3000 s, the logic error overlaps with the human misunderstanding, making the risk to further fluctuate.

Such a real-time risk figure displays how interpretation conflict behaves in different situations. When the interpretation conflict risk appears high, it is time to consider whether an interpretation conflict has occurred rather than always a fault or failure. Operators are thus better able to take more targeted measures to resolve the conflict (Wen et al., 2022). Typically, the violent fluctuations are more likely to be the

superposition of observation conflict and interpretation conflict, such as Situation 4 in 2001-2500 s; whereas a lower risk may indicate common measurement errors, e.g., Situation 1 in 501-1500 s.

5.5. Discussion

From the above sections, the following key points can be emphasized and discussed.

Noise effect on observation and observation conflict. Severe observation conflict may occur when the observations deviate from the setpoint significantly. Additionally, the VOD is clearly beyond the limit once a noise is introduced, including measurement error, sensor fault, logic error, and human mistake and misunderstanding. This is common in process operations, and it implies that real-time monitoring and response are essential.

Difficulty to identify and assess interpretation conflict. From the human response perspective, interpretation conflict is expressed as confusion, like mode confusion and automation surprise. As the operators can only judge and interpret from observations, the observations cannot indicate the interpretation conflict alone. This confirms that the logic errors and cyberattacks on the logic solver or AI model are usually hidden and invisible. On the other hand, from the risk assessment results of the time step 2001 s, an interpretation conflict is instant once the logic error happens, and the risk reaches high sharply. Therefore, it is necessary to use risk-based approaches to predict and assess it.

Noise effect on interpretation conflict. In 2001-3000 s, the logic error happens. Once the interpretation conflict occurs, it is easy for the operator to misunderstand the

situation and take the wrong action. Such noise may have different types and forms; traditionally, it may be the mechanical problem or programming problem of the logic solver. Any other interference or impairment of the computing capability, for example, DoS attacks, might have a similar interpretation conflict. Usually, data pollution, insufficient data volume, and limited training can degrade AI's applicability, integrity, and robustness. Consequently, they may force the AI to interpret incorrectly, which needs further verification.

Bounded noise or unbounded noise. The traditional control theory solves the disturbance of bounded noise well. However, the noises caused by sensor faults, cyberattacks, and human errors are usually unbounded, especially, from the security perspective. These noises may have similar fluctuations in the observations. This study proposes the triangular distribution to set the noise boundaries. From the time series, the observations fluctuate and hide the interpretation conflict. It can be challenging for inexperienced operators to judge. It also confirms that there are undetectable logic problems or the hacker is reluctant to be detected with apparent abnormalities.

Distance to measure interpretation conflict. Measuring the distance of probability vectors between humans and AI to measure the interpretation conflict is the most challenging part of this study since interpretation is intangible. It refers to the techniques in deep learning, which usually use Softmax to obtain the probability vector and cross-entropy to measure the loss. Compared to the Manhattan and Euclidean distances, cross-entropy-based distance measurement is suitable for interpretation conflict assessment.

Resistance of advanced control and data-driven control to interpretation conflict. This study employs s a linear model-based control (i.e., PID) on a classic model with a single input and output to show how various noises generate interpretation conflicts. The reason for choosing PID instead of more advanced or even data-driven control is that PID is still the primary choice in process industries. One hypothesis is that advanced control (e.g., model predictive control) or AI control might counteract or respond differently to interpretation conflicts. Especially for the time series data, recurrent neural networks (RNN) and their variants can be suitable to buffer the disturbances. In the meantime, performance indicators of AI models can evaluate the noise effect and may contribute to estimating the conflict, which needs further study. Eventually, if the noise/disturbance can be suppressed, it may not trigger human-AI conflict. On the other hand, AI algorithms usually display the black-box issue; therefore, combining physical model-based control and data-based control may produce better performance, yet challenging.

5.6. Conclusions

This study deconstructs the cognitive processes of humans and AI by proposing the concept of interpretation conflict, extending the situation awareness to interpretation conflict, and proposing the methodology to identify the situations of interpretation conflict, further evaluating its probability, and quantifying its severity and risk. The proposed methodology has been applied to a two-phase separator unit. The simulation shows when interpretation conflict occurs, the observations are quite similar to

traditional faults. Significant observation conflict triggers interpretation conflict. Also, various noises can cause interpretation conflict, including sensor faults, logic errors, cyberattacks, human mistakes, and misunderstandings. When there is an interpretation conflict, humans may not take the right action timely, allowing a conflict to lead to catastrophic consequences.

This paper emphasizes the need for assessing interpretation conflict to discover the difference between intelligence control and human-centric control to optimize the controller design from a safety perspective. Considering interpretation conflict as unbounded noise provides a broader idea for model predictive control and other data-driven control design. As intelligent machines approach full automation, situation awareness becomes critical. Incorporating this in design and operation will help achieve safer and more robust processes. This study does not consider multiple inputs and multiple outputs. This is an essential aspect of AI and how humans will consider multi-parameters data (sensor data fusion) differently. This is a future research direction.

Chapter 6: An Integration Study

6.1. Introduction

The targets of cyber threats have been ever-changing with the technology evolution. In the 1980 s, cyberattacks were mostly on individual computers, then network computers or corporate systems. However, industrial systems and connected devices now show more value and attraction to attackers. One research shows that industrial control system (ICS) vulnerabilities have kept increasing annually, 90% of the reported ICS vulnerabilities have a low attack complexity, and 71% are remotely exploitable (Claroty, 2021). Nevertheless, the attack tactics and techniques on ICS are also diversifying and complex. Process industries, for example, oil & gas, chemical, energy, nuclear, and water/wastewater, present more incidents and severe impairment based on professional reports (European Union Agency for Cybersecurity [ENISA], 2022; Kaspersky, 2022) and review papers (Iaiani, Tugnoli, Bonvicini, et al., 2021a). Notable cyber incidents on ICS are listed in Table 6.1.

Table 6.1: Cyber incidents on ICS.

Year	Type	Incident
2000	Attack	Attack on Maroochy Water Plant
2008	Attack	Turkey Pipeline Explosion
2010	Malware	Stuxnet
2011	Malware	Duqu/Flame/Gauss
2012	Campaign	Gas Pipeline Cyber Intrusion Campaign
2012	Malware	Shamoon
2013	Malware	Havex
2014	Attack	German Steel Mill
2014	Malware	Black Energy
2014	Campaign	Berserk Bear No. 1
2015	Attack	Ukraine Power Grid Attack No. 1
2016	Attack	Kemuri water company
2016	Malware	Return of Sharnoon
2016	Attack	Ukraine Power Grid Attack No. 2
2017	Malware	Crash Override
2017	Group	Advanced Persistent Threat 33
2017	Attack	NotPetya
2017	Campaign	Berserk Bear No. 2
2017	Malware	Triton
2021	Attack	Oldsmar Water Plant Attack

These incidents demonstrate that cyberattacks manifest diverse techniques to provoke various human-machine conflicts. Moreover, they are also sabotaging and dedicating efforts to overrun the system. This thesis is motivated by three notable cyber incidents (Table 6.2). Therefore, this chapter conducts an integrated study to simulate the process of these incidents and associated conflicts, then illustrates the strategies by citing the risk model in the previous three chapters.

Table 6.2: Cyber incidents and associated conflicts.

Incident	Key technique	Conflict	Reference
Water plant poisoning attack	Setpoint modification: ● Increase sodium hydroxide levels from 100ppm to 11,100ppm	Observation conflict	Chapter 3
Stuxnet attack on centrifuges	FDI: ● Modify the frequency of the motors to 1,410 Hz and then to 2 Hz, and then to 1,064 Hz	Interpretation conflict	Chapter 5
Black Energy attack on power grids	DoS: ● Disable the control system	Action conflict	Chapter 4

Chapters 3 and 5 use two-phase separators with PID controller to verify the observation conflict and interpretation conflict corresponding to sensor fault and logic error, respectively. This chapter affirms the conflict scenarios due to cyberattacks on the same two-phase separator. The simulation presents how to help operators distinguish whether it is a fault or an attack and clarifies how to identify conflicts and evaluate conflict risks.

6.2. Methodology

This integration study consists of attack simulation and conflict risk management (Figure 6.1). Each attack corresponds with one type of conflict, and the procedures of conflict risk management are conducted sequentially.

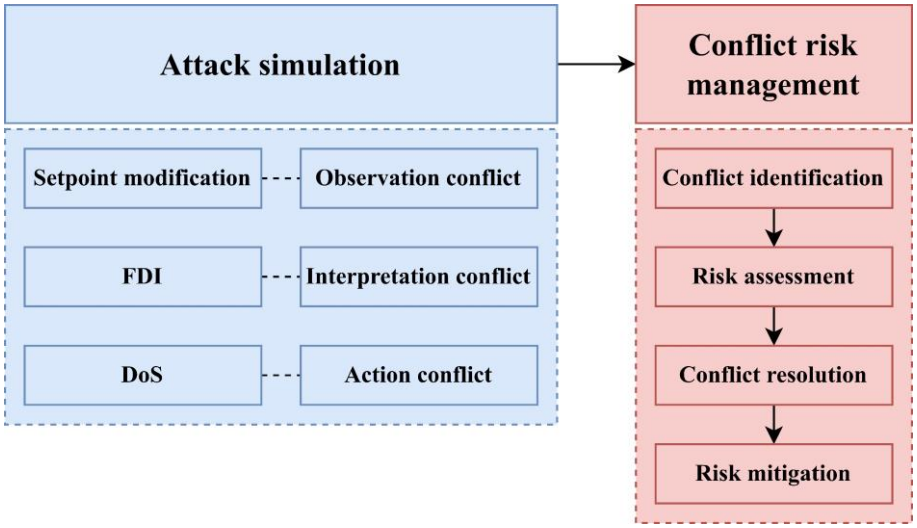

Figure 6.1: Research flowchart.

This integration study applies the same two-phase separator in chapters 3 and 5 with unified assumptions and conditions. The attacks are shown in Table 6.3, assuming the attack occurs in 501-1000 s and repeating the simulation for other attacks after reset. Three attacks are simulated: setpoint modification, FDI, and DoS. The simulation model established by MATLAB/Simulink R2021a is shown in Figure 6.2. A constant is introduced to simulate the setpoint modification attack; a generated date sheet with a triangular distribution is employed to simulate the FDI attack; a plus-minus module is referenced to indicate that the input is 0, which simulates the DoS attack.

Table 6.3: Attacks on the two-phase separator.

Time	Attack	Simulation	Expression
0-500 s	-	-	-
501-1000 s	Setpoint modification	Change the setpoint from 0.5 to 0.8	$r = 0.8$
	FDI	Manipulate the feedback sensor observation with a triangular distribution [0.2, 0.7, 0.9]	$x = [0.2,0.7,0.9]$
	DoS	Change the input with 0	$u = 0$

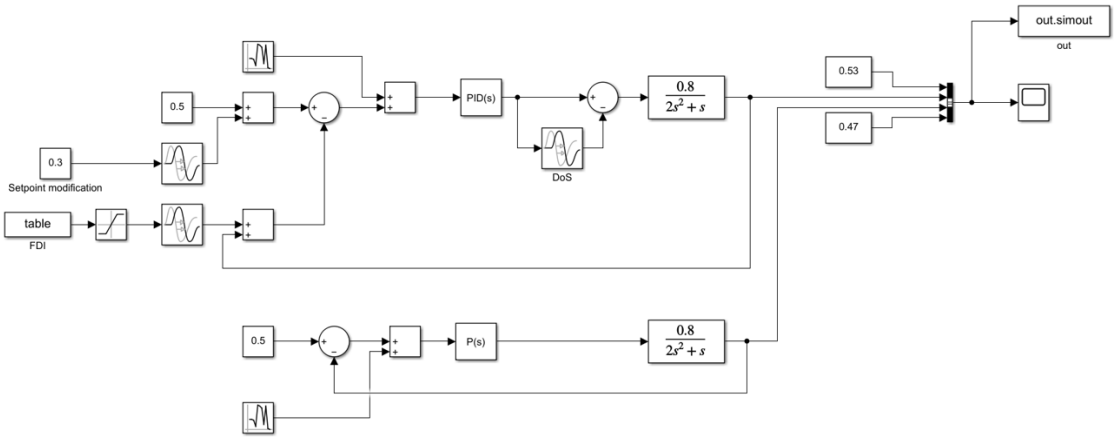

Figure 6.2: The simulation model.

6.3. Results and discussion

6.3.1. Conflict identification and risk assessment

6.3.1.1. Setpoint modification

The observations of the setpoint modification attack are shown in Figure 6.3. Setting manual control as the reference, once the hacker modifies the setpoint, the PID controller has an instant response with little resistance (compared with the MPC resistance in Chapter 4). The process value rapidly stabilizes around the new setpoint. Apparently, it occurs as an observation conflict ($VOD \gg 0.01$).

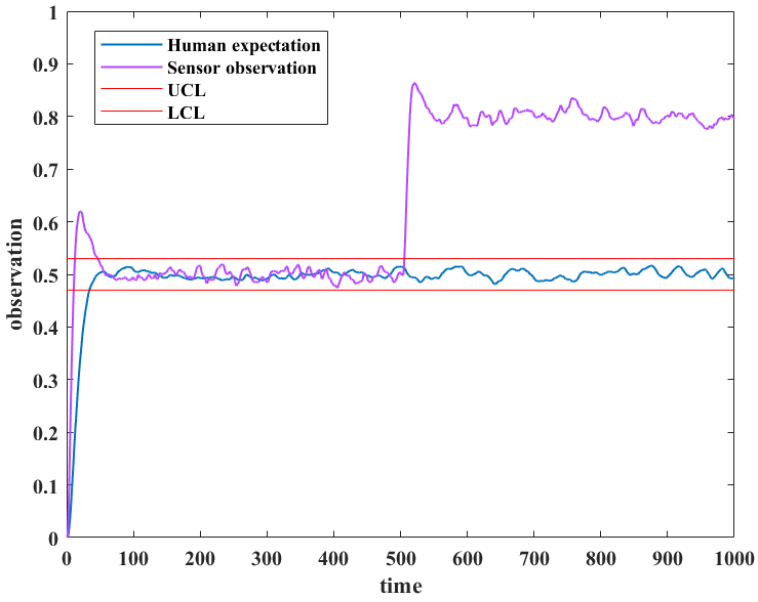

Figure 6.3: Attack results of setpoint modification.

Referring to the risk model in Chapter 3, the probability, severity, and risk of this observation conflict are $P \approx 1$, $S \approx 1$, and $R \approx 1$. The max severity is 1. Thus, the risk reaches the maximized value.

In the attack of setpoint modification, it is easy for the operator to notice the abnormal situation. For example, in the incident of the water plant attack, the operator has already discovered that someone is changing the setpoint value on the HMI without even checking the actual field level. Therefore, this is what complex attacks want to avoid, and hackers will take covert measures to cover up the modification of the process value, at least the displayed value on the HMI.

Furthermore, if the sensor occurs a bias fault, it may also show similar stable deviation. As we advocate in our follow-up research, fault resolution can be attempted first, and if it proves invalid, attack resolution should be considered immediately.

6.3.1.2. FDI

To avoid apparent abnormalities being spotted by operators, hackers often fake normal values and manipulate sensor observations of the feedback loop. The Stuxnet attack exploited several vulnerabilities and applied comprehensive attack techniques, and FDI was one of the key tactics. Figure 6.4 shows how the hacker presents the regular false sensor observation until the operator notices the actual liquid level has been relatively low. It shows automation surprise, a kind of interpretation conflict triggered by observation conflict, which is Situation 3, referring to Chapter 5.

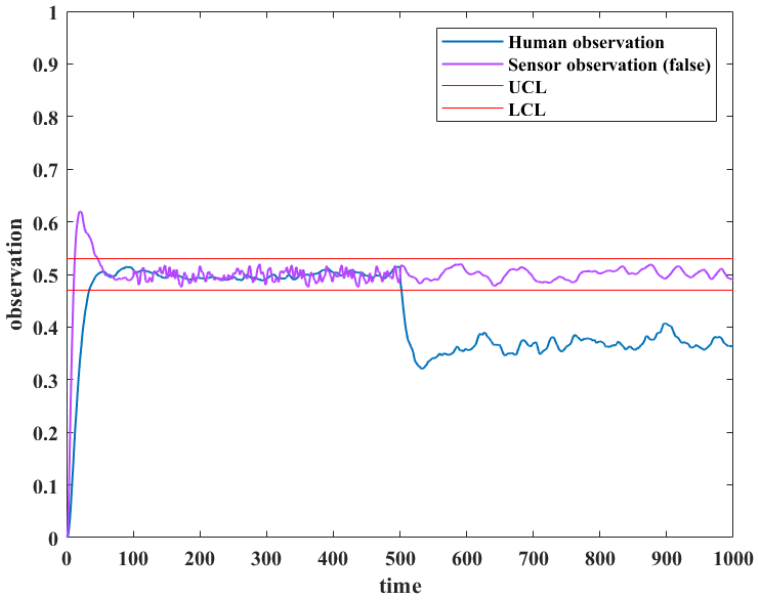

Figure 6.4: Results of FDI attack.

Referring to the risk model in Chapter 5, the interpretation conflict risk is calculated and shown in Figure 6.5. The risk is around 1.52 in 501-1000 s, noticeably higher than the threshold of 0.5 and higher than the half-maximum risk (1.34).

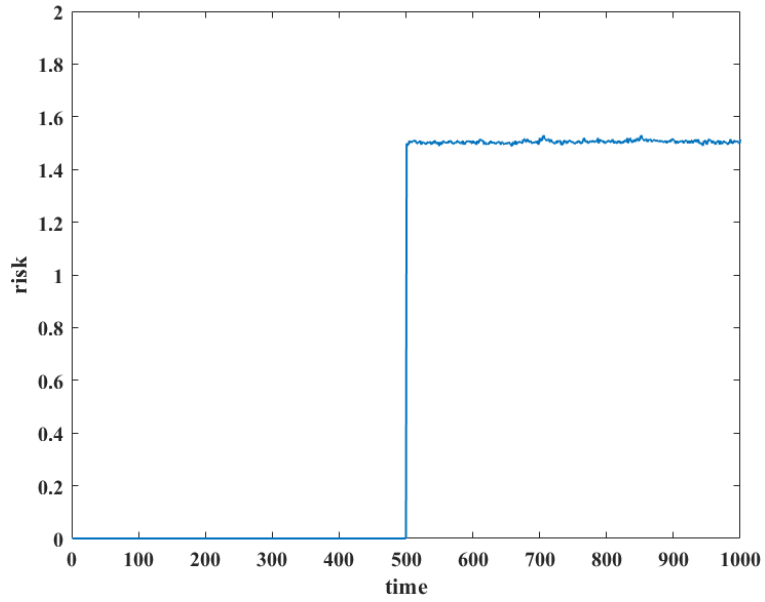

Figure 6.5: Risk of FDI attack.

Based on traditional reliability engineering, this situation is likely caused by sensor malfunction or controller malfunction. However, fault resolution to the sensor or controller cannot resolve the conflict. In Chapter 4, we have proposed a criterion (half maximum risk) to judge whether it is a fault or an attack and when to switch the resolution strategy. In 501-1000 s, the conflict risk exceeds half of the maximum risk, it should be considered that this is not a fault but an attack.

In addition, once interpretation conflicts arise, it is fundamental to ensure that the system can switch to manual control mode, although the priority of the two has always been debated.

6.3.1.3. DoS

DoS is the top cyberattack, compared with all other techniques. When DoS attacks the process control system, a typical phenomenon is that the value of the input variable (manipulated variable) returns to 0, forcing the controller to lose the capability to adjust

and resist the disturbance. Figure 6.6 displays the consequence of DoS on the PID controller. If the system cannot enter manual mode, this will present an apparent action conflict.

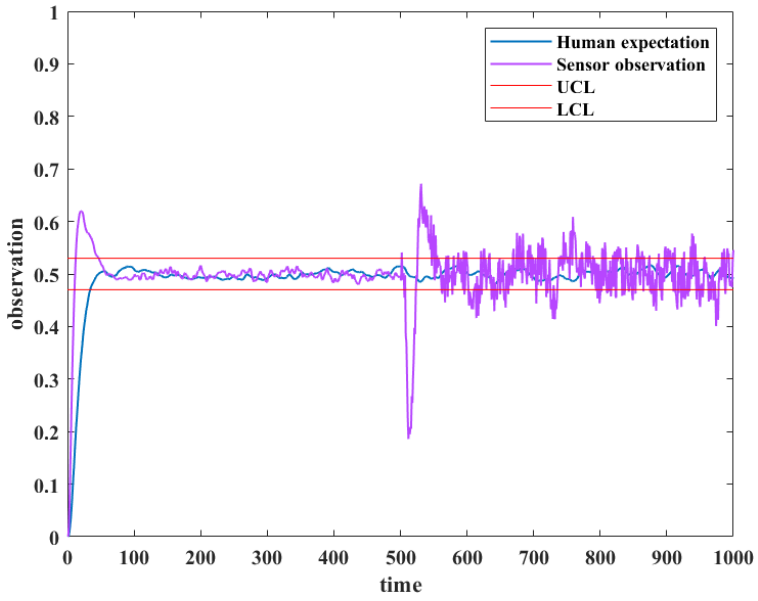

Figure 6.6: Results of DoS attack.

Referring to the risk model in Chapter 4, the probability, severity, and risk of the action conflict are $P = 1$, $S = 1$, and $R = 1$. The maximum risk is 1 here. It proves that when there is no input from the controller to the dynamic system, the system enters into a state of disorder with the highest risk. In this case, to be able to switch to manual control will be highly critical.

This situation is phenomenologically equivalent to a controller malfunction or logic error since the value of the input variable is 0, and the operator can quickly discover this abnormality from the HMI. However, it is vital to alert and train the operators about the common sense of DoS attacks, to take immediate measures to ensure the manual operation of the system.

6.3.2. Conflict resolution

The occurrence of conflicts is often sudden; hence, conflict resolution requires emergency measures and responses. Based on the previous analysis in Chapter 3, the essence of conflict is still a fault. Therefore, conflict resolution usually depends on fault resolution. Considering cyberattacks, this study divides conflict resolution into two paths: fault resolution and attack resolution. Chapter 3 has listed common methods of fault resolution. Here it will focus on the solution to cyberattacks. Chapter 4 has developed the strategy for switching from fault resolution to attack resolution.

Once the process system is attacked, the most critical task is to ensure the stability of the main chemical process, which may require isolating the attacked subsystem. For the attack of setpoint modification, FDI, or DoS, the immediate measure could be to restrict access from the other party's IP address, further close the corresponding propagation port, or even disconnect the corresponding switch port to shut down the network. It is unwise to shut down the system directly or disconnect the network, as this already constitutes a mishap or incident.

From a long-term perspective, precautions should be conducted ahead, for example, virus scanning, code audit, penetration detection, black and white box test, and red team assessment. It requires the cooperation of process engineers and IT engineers. Nevertheless, all process engineers and operators should be educated about the presence and possibility of cyberattacks. It is their responsibility and duty to identify and judge the conflict due to attacks or faults.

6.3.3. Risk mitigation

The mitigation of conflict risk due to cyberattacks usually depends on IT solutions. From the safety perspective, the mitigation procedures of conflict risk could follow the classic techniques in safety engineering, for example, the Hierarchy of Controls (*Hierarchy of Controls | NIOSH | CDC*, n.d.): elimination, substitution, engineering control, administrative control, and personal protective equipment (PPE).

First, the ideal solution to conflict risk is full automation to avoid the cooperative scenario with humans and shield from human intervention. Admittedly, it relies on the inherent safety design of the machine and still faces cyber threats and intentional sabotage, which have been simulated in this chapter.

Since full automation is far from its realization, humans will still be there for substitution. Therefore, mutually responsive collaboration and human-centered AI can be advocated to strengthen the capability of situation awareness of AI, reserve the priority of humans, and reduce human error probability.

At present, a more practical way is still to apply engineering controls to the process systems, for example, the fail-safe design, fault-tolerant control, and perhaps even conflict-tolerant control. More robust and reliable control could prevent foreseeable conflicts and relieve hassles for operators. It is achievable in the digitalization age and should be propagated in system design.

Besides, administrative controls and management approaches are vital since this requires the cooperation of process engineers and IT engineers for digitalization or

cybersecurity. Although this is a reactive response, an accurate and rapid response can minimize the conflict risk.

The last is how to protect humans in an emergency, like PPE as the last defense. It requires to think how to escape from a conflict or how humans abandon a conflict with machines and self-rescue. In essence, the current human-machine conflict results from the two parties. The withdrawal of any party will eliminate such conflict. Humans can make more rational choices, as the last line of defense, to abandon the conflict is also a kind of self-protection.

6.4. Conclusions

The diversity of cyberattacks highlights the innumerable possibilities of sabotage and triggering various conflict scenarios. It also presents that cyberattacks could impact the whole loop of recognition and action, leading to observation, interpretation, and action conflicts.

For process system engineering, cyberattacks mimic some phenomena of common faults and failures. It is a reminder that cyberattacks could be everywhere and at any time. Real-time conflict risk would monitor cyberattacks, and the rule of half maximum risk could be considered a principle to switch resolution strategy for a fault to that of an attack.

Furthermore, the priority and authority of human control should be maintained before full automation, especially since it should be easy to hand over to humans in attack scenarios, avoiding being hijacked in the long term. Human choices remain the most

reliable until machines have reasonable value judgments.

Chapter 7: Conclusions and Recommendations

7.1. Conclusions

This thesis conducts five research tasks corresponding to Chapters 2 to 6. Chapter 2 is a systematic review of the literature on human-machine conflict and myths of misusing data models; Chapter 3 is the study of conflict due to sensor fault; Chapter 4 is about conflict due to cyberattacks; Chapter 5 is the interpretation conflict from situation awareness; and Chapter 6 is an integration study considering a loop of recognition and action in cyberattack scenarios.

This thesis defines and classifies human-machine conflict, explores the causes and evolution of conflict, and how to express conflict mathematically, especially in conditions of sensor faults, cyberattacks, human errors, and sabotage. Subsequently, evaluation models of conflict risk are proposed, such as identifying conflicts and evaluating the probability and severity of conflicts. By utilizing the risk models, application and case studies are carried out on a two-phase separator and a CSTR model. The thesis draws valuable conclusions from mathematical derivation, modeling, and case study results.

7.1.1. The concept of human-machine conflict

This thesis verifies that process automation and digitization will bring new types of risks, of which human-machine conflict is worthy of attention and vigilance. It appears as machines become increasingly intelligent, automated, and digitalized, presenting a challenge to humans. Human-machine conflict is often misunderstood as a traditional

fault, malfunction, or failure of the machine; however, such a view does not consider the existence of humans and the interaction between the two intelligent agents. Therefore, human-machine conflict is a new subject matter in digitalized age.

7.1.2. The nature of human-machine conflict

This thesis defines and discusses the definition and classification of observation conflict, interpretation conflict, and action conflict. Observation conflict generally shows numerical differences, while interpretation conflict involves value judgments, and action conflict is the product of the evolution of the former two. This thesis provides fundamental mathematical expressions of human-machine conflict to reveal how conflicts evolve, diverge, converge, and resolve.

From the phenomenon, the action conflict may lead to the ultimate risk. However, as the system's executive components become more unified, humans and AI will act on the same actuator, making observation and interpretation conflict particularly critical.

7.1.3. Causes of human-machine conflict

Various reasons can lead to human-machine conflict, including sensor faults, logic errors, cyberattacks, human mistakes, misunderstandings, and even sabotage. The thesis demonstrates sensor faults and cyberattacks leading to human-machine conflict. This will be a significant obstacle to further application of digitalization and poses challenges to reliability engineers and IT engineers. A strong collaborative approach needs to be taken in order to mitigate the conflict risk.

7.1.4. Risk-based monitoring of human-machine conflict

This thesis proposes a scheme to measure the probability, severity, and risk of conflict by distance, which is precisely to measure the differences and gaps between humans and machines. Since conflicts may happen suddenly, it leaves little time to identify and respond. This emphasizes the importance of real-time risk monitoring. It is the value that the risk models are proposed. On this basis, conflict resolution and risk mitigation can be conducted.

7.2. Recommendations

This thesis has attempted to introduce the new concept of human-machine conflict, yet opportunities still exist that could be further studied. Since it is difficult to quantify the observation differences between humans and machines, this thesis has to choose relatively simple models, such as the liquid level control model, to obtain a univariate comparison. No doubt that multivariate situations should be considered, and sensor fusion or data fusion should be compared with human recognition. Moreover, this thesis linearizes some models and process systems to simplify the problems, which are nonlinear in practice. Though it is a solution to reach significant outputs, it deserves further research and improvement in future study. Nevertheless, as the initial research of human-machine conflict, such models are illustrative.

Also, AI is still in the narrow or weak AI stage, and its intelligence level has not surpassed that of human beings. This difference in intelligence is the primary source of human-machine conflict. Given that the application of AI in the process industry is far

less than in aviation and autonomous vehicles, the human-machine conflict studied in this thesis needs to be empirically studied in multiple sectors.

In future, when the intelligence level of AI surpasses that of human beings, the human-machine conflict will take another form or may cause more severe crises. Therefore, the design concept of human-centered AI should be emphasized. The topic of human-machine conflict needs further attention and exploration in academic and industrial practice.

References

Aaltonen, I., Salmi, T., & Marstio, I. (2018). Refining levels of collaboration to support the design and evaluation of human-robot interaction in the manufacturing industry. *Procedia CIRP*, *72*, 93–98. https://doi.org/10.1016/j.procir.2018.03.214

Acioli, C., Scavarda, A., & Reis, A. (2021). Applying industry 4.0 technologies in the COVID–19 sustainable chains. *International Journal of Productivity and Performance Management*, *70*(5), 988–1016. https://doi.org/10.1108/IJPPM-03-2020-0137

Ahmed, S. (2016). Identification from step response-The integral equation approach. *Canadian Journal of Chemical Engineering*, *94*(12), 2243–2256. https://doi.org/10.1002/cjce.22645

Ahmed, S. (2021). Artificial intelligence and machine learning for process safety: Points to ponder. *Process Safety Progress*, *40*(4), 189–190. https://doi.org/10.1002/prs.12321

Ahmed, S., & Imtiaz, S. A. (2015). Identification of MIMO continuous-time models using simultaneous step inputs. *Industrial and Engineering Chemistry Research*, *54*(29), 7251–7260. https://doi.org/10.1021/acs.iecr.5b00481

Alhaji, B., Beecken, J., Ehlers, R., Gertheiss, J., Merz, F., Müller, J. P., Prilla, M., Rausch, A., Reinhardt, A., Reinhardt, D., Rembe, C., Rohweder, N. O., Schwindt, C., Westphal, S., & Zimmermann, J. (2020). Engineering human–

machine teams for trusted collaboration. *Big Data and Cognitive Computing*, *4*(4), 1–30. https://doi.org/10.3390/bdcc4040035

Aptiv, Audi, Baidu, BMW, Continental, FCA, Here, Infineon, Intel, & Volkswagen. (2019). Safety first for automated driving 2019. In *White paper of different car manufacutres and suppliers.* https://www.press.bmwgroup.com/global/article/attachment/T0298103EN/434404

Arauz, T., Chanfreut, P., & Maestre, J. M. (2022). Cyber-security in networked and distributed model predictive control. *Annual Reviews in Control*, *53*(October 2021), 338–355. https://doi.org/10.1016/j.arcontrol.2021.10.005

Arunthavanathan, R., Khan, F., Ahmed, S., Imtiaz, S., & Rusli, R. (2020). Fault detection and diagnosis in process system using artificial intelligence-based cognitive technique. *Computers & Chemical Engineering*, *134*, 106697. https://doi.org/https://doi.org/10.1016/j.compchemeng.2019.106697

Baybutt, P. (2002). Layers of protection analysis for human factors (LOPA-HF). *Process Safety Progress*, *21*(2), 119–129. https://doi.org/10.1002/prs.680210208

Bécue, A., Praça, I., & Gama, J. (2021). Artificial intelligence, cyber-threats and Industry 4.0: challenges and opportunities. In *Artificial Intelligence Review* (Vol. 54, Issue 5, pp. 3849–3886). https://doi.org/10.1007/s10462-020-09942-2

Bemporad, A., Ricker, N. L., & Morari, M. (2021). *Model predictive control toolbox getting started guide*. The MathWorks, Inc.

https://www.mathworks.com/help/pdf_doc/mpc/mpc_gs.pdf

Benson, C., Argyropoulos, C. D., Dimopoulos, C., Mikellidou, C. V., & Boustras, G. (2021). Safety and risk analysis in digitalized process operations warning of possible deviating conditions in the process environment. *Process Safety and Environmental Protection*, *149*, 750–757. https://doi.org/10.1016/j.psep.2021.02.039

Bequette, B. W. (2019). 110th anniversary: Commentary: The smart human in smart manufacturing. *Industrial and Engineering Chemistry Research*, *58*(42), 19317–19321. https://doi.org/10.1021/acs.iecr.9b03544

Beringer, D. B., & Harris, H. C. (1999). Automation in general aviation: Two studies of pilot responses to autopilot malfunctions. *International Journal of Aviation Psychology*, *9*(2), 155–174. https://doi.org/10.1207/s15327108ijap0902_5

Bhavsar, P., Srinivasan, B., & Srinivasan, R. (2017). Quantifying situation awareness of control room operators using eye-gaze behavior. *Computers and Chemical Engineering*, *106*, 191–201. https://doi.org/10.1016/j.compchemeng.2017.06.004

Boy, G. A. (2017). *The handbook of human-machine interaction: a human-centered design approach*. CRC Press.

Bredereke, J., & Lankenau, A. (2005). Safety-relevant mode confusions - Modelling and reducing them. *Reliability Engineering and System Safety*, *88*(3), 229–245. https://doi.org/10.1016/j.ress.2004.07.020

Briken, K. (2020). Welcome in the machine: Human-machine relations and

knowledge capture. *Capital and Class*, *44*(2), 159–171. https://doi.org/10.1177/0309816819899418

Buncefield Major Incident Investigation Board. (2008). *The Buncefield Incident 11 December 2005* (Vol. 2, Issue December).

Bundesamt für Sicherheit in der Informationstechnik [BSI]. (2022). *Industrial control system security: Top 10 threats and countermeasures 2022*. https://www.allianz-fuer-cybersicherheit.de/SharedDocs/Downloads/Webs/ACS/DE/BSI-CS/BSI-CS_005E.html

Bzdok, D., Altman, N., & Krzywinski, M. (2018). Statistics versus machine learning. *Nature Methods*, *15*(4), 233–234. https://doi.org/10.1038/nmeth.4642

Campo-Flores, A. (2021, February 9). U.S. news: Hacker tampers with Florida water system. *Wall Street Journal*. https://qe2a-proxy.mun.ca/login?url?url=https://www.proquest.com/newspapers/u-s-news-hacker-tampers-with-florida-water-system/docview/2487322588/se-2?accountid=12378

Canonico, L. B. (2019). *Human-machine teamwork: An exploration of multi-agent systems, team cognition, and collective intelligence.* [Clemson University]. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=psyc17&NEWS=N&AN=2020-28118-098

Castelfranchi, C. (2000). Conflict ontology. In *Computational conflicts* (pp. 21–40). Springer.

Chen, T. (2010). Stuxnet, the real start of cyber warfare? *IEEE Network*, *24*(6), 2–3. https://doi.org/10.1109/MNET.2010.5634434

Claroty. (2021). *Claroty biannual ICS risk & vulnerability report:2H 2021*. https://security.claroty.com/biannual-ics-risk-vulnerability-report-2H-2020

Cohen, Y., & Singer, G. (2021). A smart process controller framework for Industry 4.0 settings. *Journal of Intelligent Manufacturing*, *32*(7), 1975–1995. https://doi.org/10.1007/s10845-021-01748-5

Combéfis, S., Giannakopoulou, D., & Pecheur, C. (2016). Automatic detection of potential automation surprises for ADEPT models. *IEEE Transactions on Human-Machine Systems*, *46*(2), 267–278. https://doi.org/10.1109/THMS.2015.2424851

Conflict | Definition of Conflict by Merriam-Webster. (n.d.). Retrieved November 4, 2021, from https://www.merriam-webster.com/dictionary/conflict

Coombs, C., Stacey, P., Kawalek, P., Simeonova, B., Becker, J., Bergener, K., Carvalho, J. Á., Fantinato, M., Garmann-Johnsen, N. F., Grimme, C., Stein, A., & Trautmann, H. (2021). What is it about humanity that we can't give away to intelligent machines? A European perspective. *International Journal of Information Management*, *58*. https://doi.org/10.1016/j.ijinfomgt.2021.102311

Damacharla, P., Javaid, A. Y., Gallimore, J. J., & Devabhaktuni, V. K. (2018). Common metrics to benchmark human-machine teams (HMT): A review. *IEEE Access*, *6*, 38637–38655. https://doi.org/10.1109/ACCESS.2018.2853560

DeFazio, P. A., & Larsen, R. (2020). Final committee report—The design, development & certification of the Boeing 737 MAX. *The US House Committee on Transportation and Infrastructure, Subcommittee on Aviation. Washington, DC: US House Committee on Transportation and Infrastructure*.

Dehais, F., Peysakhovich, V., Scannella, S., Fongue, J., & Gateau, T. (2015). "Automation surprise" in aviation: Real-time solutions. *Conference on Human Factors in Computing Systems - Proceedings*, *2015-April*(April), 2525–2534. https://doi.org/10.1145/2702123.2702521

Dehais, F., Tessier, C., & Chaudron, L. (2003). GHOST: Experimenting conflicts countermeasures in the pilot's activity. *IJCAI International Joint Conference on Artificial Intelligence*, 163–168.

Deutsch, M. (1990). Sixty years of conflict. *International Journal of Conflict Management*, *1*(3), 237–263. https://doi.org/10.1108/eb022682

Dowell, A. M. (1998). Layer of protection analysis for determining safety integrity level. *ISA Transactions*, *37*(3), 155–165. https://doi.org/10.1016/s0019-0578(98)00018-4

Durand, H., & Wegener, M. (2020). Mitigating safety concerns and profit/production losses for chemical process control systems under cyberattacks via design/control methods. *Mathematics*, *8*(4). https://doi.org/10.3390/math8040499

Easterbrook, S. (1991). Handling conflict between domain descriptions with computer-supported negotiation. *Knowledge Acquisition*, *3*(3), 255–289.

https://doi.org/10.1016/1042-8143(91)90007-A

Easterbrook, S. M. (1994). Resolving requirements conflicts with computer-supported negotiation. *Requirements Engineering: Social and Technical Issues*, 41–65.

Easterbrook, S. M., Beck, E. E., Goodlet, J. S., Plowman, L., Sharples, M., & Wood, C. C. (1993). *A survey of empirical studies of conflict* (Issue December 2014). https://doi.org/10.1007/978-1-4471-1981-4_1

Edwards, E., & Lees, F. P. (1971). The development of the role of the human operator in process control. *IFAC Proceedings Volumes*, *4*(3), 138–144. https://doi.org/10.1016/s1474-6670(17)68589-6

Ehie, I. C., & Chilton, M. A. (2020). Understanding the influence of IT/OT convergence on the adoption of Internet of Things (IoT) in manufacturing organizations: An empirical investigation. *Computers in Industry*, *115*, 103166. https://doi.org/10.1016/j.compind.2019.103166

Endsley, M. R. (1988a). Design and evaluation for situation awareness enhancement. *Proceedings of the Human Factors Society Annual Meeting*, *32*(2), 97–101.

Endsley, M. R. (1988b). Situation awareness global assessment technique (SAGAT). *Proceedings of the IEEE 1988 National Aerospace and Electronics Conference*, 789–795.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors*, *37*(1), 32–64. https://doi.org/10.4324/9781315092898-13

Endsley, M. R., & Kaber, D. B. (1999). Level of automation effects on performance,

situation awareness and workload in a dynamic control task. *Ergonomics*, *42*(3), 462–492. https://doi.org/10.1080/001401399185595

European Agency for Safety & Health at Work [EU-OSHA]. (2018). *Foresight on new and emerging occupational safety and health risks associated with information and communication technologies and work location by 2025*. https://osha.europa.eu/pt/publications/foresight-new-and-emerging-occupational-safety-and-health-risks-associated

European Union Agency for Cybersecurity [ENISA]. (2022). ENISA threat landscape 2022. In *European Union Agency for Cybersecurity* (Issue October). https://www.enisa.europa.eu/publications/enisa-threat-landscape-2022

Ferreira, C., Figueira, G., & Amorim, P. (2021). Scheduling human-robot teams in collaborative working cells. *International Journal of Production Economics*, *235*(May 2020), 108094. https://doi.org/10.1016/j.ijpe.2021.108094

Flemisch, F., Abbink, D. A., Itoh, M., Pacaux-Lemoine, M. P., & Weßel, G. (2019). Joining the blunt and the pointy end of the spear: towards a common framework of joint action, human–machine cooperation, cooperative guidance and control, shared, traded and supervisory control. *Cognition, Technology and Work*, *21*(4), 555–568. https://doi.org/10.1007/s10111-019-00576-1

Freschi, C., Ferrari, V., Melfi, F., Ferrari, M., Mosca, F., & Cuschieri1, A. (2013). Technical review of the da Vinci surgical telemanipulator. *The International Journal of Medical Robotics and Computer Assisted Surgery*, *9*, 396–406.

https://doi.org/10.1002/rcs.1468

Frohm, J., Lindström, V., Stahre, J., & Winroth, M. (2008). Levels of automation in manufacturing. *Ergonomia - an International Journal of Ergonomics and Human Factors*, *30*(3).

Fryman, J., & Matthias, B. (2012). Safety of industrial robots: From conventional to collaborative applications. *7th German Conference on Robotics, ROBOTIK 2012*, 51–55.

Garimella, P. K. (2018). IT-OT integration challenges in utilities. *Proceedings on 2018 IEEE 3rd International Conference on Computing, Communication and Security, ICCCS 2018*, 199–204. https://doi.org/10.1109/CCCS.2018.8586807

Ghosh, S., & Wayne Bequette, B. (2020). Process systems engineering and the human-in-the-loop: The smart control room. *Industrial and Engineering Chemistry Research*, *59*(6), 2422–2429. https://doi.org/10.1021/acs.iecr.9b04739

Giraldo, J., Urbina, D., Cardenas, A., Valente, J., Faisal, M., Ruths, J., Tippenhauer, N. O., Sandberg, H., & Candell, R. (2018). A survey of physics-based attack detection in cyber-physical systems. *ACM Computing Surveys*, *51*(4). https://doi.org/10.1145/3203245

Gobbo, J. A., Busso, C. M., Gobbo, S. C. O., & Carreão, H. (2018). Making the links among environmental protection, process safety, and industry 4.0. *Process Safety and Environmental Protection*, *117*, 372–382. https://doi.org/10.1016/j.psep.2018.05.017

Gong, J., You, F., Wang, J. M., & Zhang, X. L. (2019). Understanding behavioural conflict between the drivers and adaptive cruise control (ACC) system in cut-in scenario. *CHIRA 2019 - Proceedings of the 3rd International Conference on Computer-Human Interaction Research and Applications*, *Chira*, 97–103. https://doi.org/10.5220/0008053600970103

Grossman, L. (2011). 2045: The year man becomes immortal. *Time (Chicago, Ill.)*, *177*(7), 1.

Hamburger, P. E. (1966). An automated method to detect potential mode confusions. *Proceedings of the 1966 21st National Conference, ACM 1966*, 321–330. https://doi.org/10.1145/800256.810711

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: concepts and techniques* (3rd ed.). Elsevier.

Heard, J., & Adams, J. A. (2019). Multi-dimensional human workload assessment for supervisory human–machine teams. *Journal of Cognitive Engineering and Decision Making*, *13*(3), 146–170. https://doi.org/10.1177/1555343419847906

Hemsley, K. E., Fisher, R. E., Kevin E. Hemsley;, & Dr. Ronald E. Fisher. (2018). History of industrial control system cyber incidents. In *Idaho National Lab* (Issue December). https://www.osti.gov/servlets/purl/1505628

Hierarchy of Controls | NIOSH | CDC. (n.d.). Retrieved December 1, 2022, from https://www.cdc.gov/niosh/topics/hierarchy/default.html

Hoc, J. M. (2000). From human – machine interaction to human – machine

cooperation. *Ergonomics*, *43*(7), 833–843. https://doi.org/10.1080/001401300409044

Hollnagel, E. (2018). *Safety-I and safety-II: the past and future of safety management*. CRC press.

Huang, C., Hang, P., Hu, Z., & Lv, C. (2021). Collision-probability-aware human-machine cooperative planning for safe automated driving. *IEEE Transactions on Vehicular Technology*, *70*(10), 9752–9763. https://doi.org/10.1109/TVT.2021.3102251

Huang, K., Di, X., Du, Q., & Chen, X. (2020). A game-theoretic framework for autonomous vehicles velocity control: bridging microscopic differential games and macroscopic mean field games. *Discrete and Continuous Dynamical Systems - Series B*, *25*(12), 4869–4903. https://doi.org/10.3934/dcdsb.2020131

Huang, Y., Chen, L., Negenborn, R. R., & van Gelder, P. H. A. J. M. (2020). A ship collision avoidance system for human-machine cooperation during collision avoidance. *Ocean Engineering*, *217*, 107913. https://doi.org/10.1016/j.oceaneng.2020.107913

Iaiani, M., Tugnoli, A., Bonvicini, S., & Cozzani, V. (2021a). Analysis of cybersecurity-related incidents in the process industry. *Reliability Engineering and System Safety*, *209*, 107485. https://doi.org/10.1016/j.ress.2021.107485

Iaiani, M., Tugnoli, A., Bonvicini, S., & Cozzani, V. (2021b). Major accidents triggered by malicious manipulations of the control system in process facilities.

Safety Science, *134*(October 2020), 105043. https://doi.org/10.1016/j.ssci.2020.105043

Iaiani, M., Tugnoli, A., Macini, P., & Cozzani, V. (2021). Outage and asset damage triggered by malicious manipulation of the control system in process plants. *Reliability Engineering and System Safety*, *213*(August 2020), 107685. https://doi.org/10.1016/j.ress.2021.107685

Inagaki, T. (2003). Automation and the cost of authority. *International Journal of Industrial Ergonomics*, *31*(3), 169–174. https://doi.org/10.1016/S0169-8141(02)00193-2

Inoue, M., Gupta, V., & Member, S. (2019). "Weak" control for human-in-the-loop systems. *IEEE Control Systems Letters*, *3*(2), 440–445.

International Ergonomics Association. (n.d.). *What is ergonomics (HFE)?* Retrieved November 3, 2022, from https://iea.cc/what-is-ergonomics/

Jost, J., Kirks, T., & Mattig, B. (2017). Multi-agent systems for decentralized control and adaptive interaction between humans and machines for industrial environments. *2017 7th IEEE International Conference on System Engineering and Technology, ICSET 2017 - Proceedings*, *October*, 95–100. https://doi.org/10.1109/ICSEngT.2017.8123427

Kaber, D. B., & Endsley, M. R. (1998). Team situation awareness for process control safety and performance. *Process Safety Progress*, *17*(1), 43–48. https://doi.org/10.1002/prs.680170110

Kaber, D. B., & Endsley, M. R. (2004). The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task. In *Theoretical Issues in Ergonomics Science* (Vol. 5, Issue 2). https://doi.org/10.1080/1463922021000054335

Kamal, S. Z., Mubarak, S. M. A., Scodova, B. D., Naik, P., Flichy, P., & Coffin, G. (2016). IT and OT convergence - Opportunities and challenges. *Society of Petroleum Engineers - SPE Intelligent Energy International Conference and Exhibition*. https://doi.org/10.2118/181087-ms

Kaspersky. (2022). Threat landscape for industrial automation systems: Statistics for H1 2022. In *AO Kaspersky Lab*. https://ics-cert.kaspersky.com/publications/reports/2021/09/09/threat-landscape-for-industrial-automation-systems-statistics-for-h1-2021/%0Ahttps://ics-cert.kaspersky.com/reports/2021/09/09/threat-landscape-for-industrial-automation-systems-statistics-for

Khan, F., Amyotte, P., & Adedigba, S. (2021). Process safety concerns in process system digitalization. *Education for Chemical Engineers*, *34*, 33–46. https://doi.org/10.1016/j.ece.2020.11.002

Klatt, K. U., & Marquardt, W. (2009). Perspectives for process systems engineering-Personal views from academia and industry. *Computers and Chemical Engineering*, *33*(3), 536–550. https://doi.org/10.1016/j.compchemeng.2008.09.002

Kolbeinsson, A., Lagerstedt, E., & Lindblom, J. (2019). Foundation for a classification of collaboration levels for human-robot cooperation in manufacturing. *Production and Manufacturing Research*, *7*(1), 448–471. https://doi.org/10.1080/21693277.2019.1645628

Kurzweil, R. (2005). *The singularity is near: When humans transcend biology*. Penguin.

Kushner, D. (2013). The real story of Stuxnet. *IEEE Spectrum*, *50*(3), 48–53. https://doi.org/10.1109/MSPEC.2013.6471059

Langner, R. (2013). To kill a centrifuge. In *The Langner Group* (Issue November). https://www.langner.com/wp-content/uploads/2017/03/to-kill-a-centrifuge.pdf

Lee, J., Cameron, I., & Hassall, M. (2019). Improving process safety: What roles for digitalization and industry 4.0? *Process Safety and Environmental Protection*, *132*, 325–339. https://doi.org/10.1016/j.psep.2019.10.021

Lee, J. H., Shin, J., & Realff, M. J. (2018). Machine learning: Overview of the recent progresses and implications for the process systems engineering field. *Computers & Chemical Engineering*, *114*, 111–121. https://doi.org/https://doi.org/10.1016/j.compchemeng.2017.10.008

Leveson, N. G., Pinnel, L. D., Sandys, S. D., Koga, S., & Reese, J. D. (1997). Analyzing software specifications for mode confusion potential. *Workshop on Human Error and System Development*, 132–146. http://www.google.co.uk/search?client=safari&rls=en-

us&q=Analyzing+software+specifications+for+mode+confusion+potential&ie=UTF-8&oe=UTF-8&redir_esc=&ei=8qwkTIeMKaiI0wTZnvy0BA

Li, J., & Vachtsevanos, G. (2014). A novel human-machine interface framework for improved system performance and conflict resolution. *PHM 2014 - Proceedings of the Annual Conference of the Prognostics and Health Management Society 2014*, 546–552.

Li, N., Oyler, D. W., Zhang, M., Yildiz, Y., Kolmanovsky, I., & Girard, A. R. (2018). Game theoretic modeling of driver and vehicle interactions for verification and validation of autonomous vehicle control systems. *IEEE Transactions on Control Systems Technology*, *26*(5), 1782–1797. https://doi.org/10.1109/TCST.2017.2723574

Li, X., & Wang, Y. (2021). Shared steering control for human-machine co-driving system with multiple factors. *Applied Mathematical Modelling*, *100*, 471–490. https://doi.org/10.1016/j.apm.2021.08.009

Li, X., Zhou, X., & Ruan, X. (2002). Conflict management in closely coupled collaborative design system. *International Journal of Computer Integrated Manufacturing*, *15*(4), 345–352. https://doi.org/10.1080/09511920210121259

Liang, G., Zhao, J., Luo, F., Weller, S. R., & Dong, Z. Y. (2017). A review of false data injection attacks against modern power systems. *IEEE Transactions on Smart Grid*, *8*(4), 1630–1638. https://doi.org/10.1109/TSG.2015.2495133

Liu, X., Zhang, J., Zhu, P., Tan, Q., & Yin, W. (2021). Quantitative cyber-physical

security analysis methodology for industrial control systems based on incomplete information Bayesian game. *Computers and Security*, *102*. https://doi.org/10.1016/j.cose.2020.102138

Lu, J., Cao, Z., Zhao, C., & Gao, F. (2019). 110th anniversary: An overview on learning-based model predictive control for batch processes. *Industrial and Engineering Chemistry Research*, *58*(37), 17164–17173. https://doi.org/10.1021/acs.iecr.9b02370

Magrini, E., Ferraguti, F., Ronga, A. J., Pini, F., De Luca, A., & Leali, F. (2020). Human-robot coexistence and interaction in open industrial cells. *Robotics and Computer-Integrated Manufacturing*, *61*(June 2018), 101846. https://doi.org/10.1016/j.rcim.2019.101846

Malle, B. F., Magar, S. T., & Scheutz, M. (2019). AI in the sky: How people morally evaluate human and machine decisions in a lethal strike dilemma. *Intelligent Systems, Control and Automation: Science and Engineering*, *95*, 111–133. https://doi.org/10.1007/978-3-030-12524-0_11

McNeese, N. J., Demir, M., Cooke, N. J., & She, M. (2021). Team situation awareness and conflict: A study of human-machine teaming. *Journal of Cognitive Engineering and Decision Making*, *15*(2–3), 83–96. https://doi.org/10.1177/15553434211017354

Mirkin, B. (2011). Data analysis, mathematical statistics, machine learning, data mining: Similarities and differences. *Computer*, 978–979.

Montgomery, D. C., & Runger, G. C. (2010). *Applied statistics and probability for engineers*. John Wiley & Sons.

Moscoso Paredes, C. T., Foss, T., & Jenssen, G. (2021). *Phantom braking in advanced driver assistance systems: Driver experience and car manufacturer warnings in owner manuals*. SINTEF.

Muhuri, P. K., Shukla, A. K., & Abraham, A. (2019). Industry 4.0: A bibliometric analysis and detailed overview. *Engineering Applications of Artificial Intelligence*, *78*(September 2018), 218–235. https://doi.org/10.1016/j.engappai.2018.11.007

Mukaidono, M., Takaoka, H., Ogihara, H., Ariyama, M., & Fujita, T. (2018). Japan's approach for the realization of the future safety concept by implementing collaborative safety technologies. *9th International Conference on Safety of Industrial Automated Systems (SIAS)*, 77–87.

Myers, P. M. (2013). Layer of protection analysis - Quantifying human performance in initiating events and independent protection layers. *Journal of Loss Prevention in the Process Industries*, *26*(3), 534–546. https://doi.org/10.1016/j.jlp.2012.07.003

Na, X., & Cole, D. J. (2015). Game-theoretic modeling of the steering interaction between a human driver and a vehicle collision avoidance controller. *IEEE Transactions on Human-Machine Systems*, *45*(1), 25–38. https://doi.org/10.1109/THMS.2014.2363124

Na, X., & Cole, D. J. (2017). Application of open-loop Stackelberg equilibrium to modeling a driver's interaction with vehicle active steering control in obstacle avoidance. *IEEE Transactions on Human-Machine Systems*, *47*(5), 673–685. https://doi.org/10.1109/THMS.2017.2700541

Naderpour, M., Lu, J., & Zhang, G. (2014). An intelligent situation awareness support system for safety-critical environments. *Decision Support Systems*, *59*(1), 325–340. https://doi.org/10.1016/j.dss.2014.01.004

Naderpour, M., Nazir, S., & Lu, J. (2015). The role of situation awareness in accidents of large-scale technological systems. *Process Safety and Environmental Protection*, *97*, 13–24. https://doi.org/10.1016/j.psep.2015.06.002

Narasimhan, S., El-Farra, N. H., & Ellis, M. J. (2022). A control-switching approach for cyberattack detection in process systems with minimal false alarms. *AIChE Journal*, *August*. https://doi.org/10.1002/aic.17875

Nazir, S., Sorensen, L. J., Overgård, K. I., & Manca, D. (2014). How distributed situation awareness influences process safety. *Chemical Engineering Transactions*, *36*, 409–414. https://doi.org/10.3303/CET1436069

Nian, R., Liu, J., & Huang, B. (2020). A review on reinforcement learning: Introduction and applications in industrial process control. *Computers and Chemical Engineering*, *139*, 106886. https://doi.org/10.1016/j.compchemeng.2020.106886

Nikmehr, N., & Moghadam, S. M. (2019). Game-theoretic cybersecurity analysis for

false data injection attack on networked microgrids. *IET Cyber-Physical Systems: Theory and Applications*, *4*(4), 365–373. https://doi.org/10.1049/iet-cps.2019.0016

Oliva, D., Hassan, S. A., & Mohamed, A. (2021). *Artificial Intelligence for COVID-19* (Vol. 358, Issue January). Springer. https://link.springer.com/10.1007/978-3-030-69744-0

Paes, R., Mazur, D. C., Venne, B. K., & Ostrzenski, J. (2020). A guide to securing industrial control networks: Integrating IT and OT systems. *IEEE Industry Applications Magazine*, *26*(2), 47–53. https://doi.org/10.1109/MIAS.2019.2943630

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans.*, *30*(3), 286–297. https://doi.org/10.1109/3468.844354

Park, J., Jeon, H., Kim, J., Kim, N., Park, S. K., Lee, S., & Lee, Y. S. (2019). Remaining and emerging issues pertaining to the human reliability analysis of domestic nuclear power plants. *Nuclear Engineering and Technology*, *51*(5), 1297–1306. https://doi.org/10.1016/j.net.2019.02.015

Patwardhan, R. S., Hamadah, H. A., Patel, K. M., Hafiz, R. H., & Al-Gwaiz, M. M. (2019). Applications of advanced analytics at saudi aramco: A practitioners' perspective. *Industrial and Engineering Chemistry Research*, *58*(26), 11338–

11351. https://doi.org/10.1021/acs.iecr.8b06205

Peres, R. S., Jia, X., Lee, J., Sun, K., Colombo, A. W., & Barata, J. (2020). Industrial artificial intelligence in Industry 4.0 -Systematic review, challenges and outlook. *IEEE Access*, 220121–220139. https://doi.org/10.1109/ACCESS.2020.3042874

Pipkorn, L., Victor, T. W., Dozza, M., & Tivesten, E. (2021). Driver conflict response during supervised automation: Do hands on wheel matter? *Transportation Research Part F: Traffic Psychology and Behaviour*, *76*, 14–25. https://doi.org/10.1016/j.trf.2020.10.001

Pistikopoulos, E. N. (2009). Perspectives in multiparametric programming and explicit model predictive control. *AIChE Journal*, *55*(8), 1918–1925. https://doi.org/https://doi.org/10.1002/aic.11965

Pistikopoulos, E. N., Barbosa-Povoa, A., Lee, J. H., Misener, R., Mitsos, A., Reklaitis, G. V., Venkatasubramanian, V., You, F., & Gani, R. (2021). Process systems engineering-The generation next? *Computers and Chemical Engineering*, *147*, 107252. https://doi.org/10.1016/j.compchemeng.2021.107252

Pizziol, S. (2013). *Prédiction des conflits dans des systèmes homme-machine*. Université de Toulouse.

Pizziol, S., Tessier, C., & Dehais, F. (2014). Petri net-based modeling of human-automation conflicts in aviation. In *Ergonomics* (Vol. 57, Issue 3, pp. 319–331). Taylor & Francis. https://doi.org/10.1080/00140139.2013.877597

Rangan, K. K., Oyama, H., & Durand, H. (2022). Actuator cyberattack handling using

Lyapunov-based economic model predictive control. *IFAC-PapersOnLine*, *55*(7), 489–494. https://doi.org/10.1016/j.ifacol.2022.07.491

Rushby, J. (2002). Using model checking to help discover mode confusions and other automation surprises. *Reliability Engineering and System Safety*, *75*(2), 167–177. https://doi.org/10.1016/S0951-8320(01)00092-8

Rüßmann, M., Lorenz, M., Gerbert, P., Waldner, M., Justus, J., Engel, P., & Harnisch, M. (2015). Industry 4.0: The future of productivity and growth in manufacturing industries. *Boston Consulting Group*, *9*(1), 54–89.

Saenz, M. J., Revilla, E., & Simón, C. (2020). Designing AI systems with human-machine teams. *MIT Sloan Management Review*, *61*(3), 1–5. https://sloanreview.mit.edu/article/designing-ai-systems-with-human-machine-teams/

Salas, E., Prince, C., Baker, D. P., & Shrestha, L. (2017). Situation awareness in team performance: Implications for measurement and training. *Situational Awareness*, *37*(1), 63–76. https://doi.org/10.4324/9781315087924-5

Salmi, T., Väätäinen, O., Malm, T., & Montonen, J. (2012). Safety challenges of transferable robotic systems. *7th German Conference on Robotics, ROBOTIK 2012*, 56–61.

Salmon, P. M., Stanton, N. A., & Jenkins, D. P. (2017). *Distributed situation awareness: Theory, measurement and application to teamwork*. CRC Press.

Sankar, G. S., & Han, K. (2020). Adaptive robust game-theoretic decision making

strategy for autonomous vehicles in highway. *IEEE Transactions on Vehicular Technology*, *69*(12), 14484–14493. https://doi.org/10.1109/TVT.2020.3041152

Sarter, N. B., & Woods, D. D. (1991). Situation awareness: A critical but ill-defined phenomenon. *The International Journal of Aviation Psychology*, *1*(1), 45–57. https://doi.org/10.4324/9781315087924-27

Shan, Y. (2021). *Human-machine interaction design for negotiation in highly automated vehicles*. Delft University of Technology.

Sharon, G., Stern, R., Felner, A., & Sturtevant, N. R. (2015). Conflict-based search for optimal multi-agent pathfinding. *Artificial Intelligence*, *219*, 40–66. https://doi.org/10.1016/j.artint.2014.11.006

Shen, J., Ye, X., & Feng, D. (2021). A game-theoretic method for resilient control design in industrial multi-agent CPSs with Markovian and coupled dynamics. *International Journal of Control*, *94*(11), 3079–3090. https://doi.org/10.1080/00207179.2020.1750707

Shneiderman, B. (2021). Human-centered AI. *Issues in Science and Technology*, *37*(2), 56–61.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489.

https://doi.org/10.1038/nature16961

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., Van Den Driessche, G., Graepel, T., & Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nature*, *550*(7676), 354–359. https://doi.org/10.1038/nature24270

Smith, K., & Hancock, P. A. (1995). Situation awareness is adaptive, externally directed consciousness. *Human Factors*, *37*(1), 137–148. https://doi.org/10.1518/001872095779049444

Sokolov, M. (2020). Decision making and risk management in biopharmaceutical engineering—opportunities in the age of covid-19 and digitalization. *Industrial and Engineering Chemistry Research*, *59*(40), 17587–17592. https://doi.org/10.1021/acs.iecr.0c02994

Stanton, N. A., Chambers, P. R. G., & Piggott, J. (2001). Situational awareness and safety. *Safety Science*, *39*(3), 189–204. https://doi.org/https://doi.org/10.1016/S0925-7535(01)00010-8

Stanton, N. A., Salmon, P. M., Walker, G. H., Salas, E., & Hancock, P. A. (2017). State-of-science: situation awareness in individuals, teams and systems. *Ergonomics*, *60*(4), 449–466. https://doi.org/10.1080/00140139.2017.1278796

Stowers, K., Brady, L. L., MacLellan, C., Wohleber, R., & Salas, E. (2021). Improving teamwork competencies in human-machine teams: Perspectives from

team science. *Frontiers in Psychology*, *12*(May), 1–6. https://doi.org/10.3389/fpsyg.2021.590290

Syfert, M., Ordys, A., Kościelny, J. M., Wnuk, P., Możaryn, J., & Kukiełka, K. (2022). Integrated approach to diagnostics of failures and cyber-attacks in industrial control systems. *Energies*, *15*(17). https://doi.org/10.3390/en15176212

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, 1–10.

Tessier, C., & Dehais, F. (2012). Authority management and conflict solving in human-machine systems. *Aerospace Lab*, *4*, 1–10.

The International Electrotechnical Commission [IEC]. (2020). *Safety in the future*. 76. https://etech.iec.ch/issue/2020-06/iec-publishes-white-paper-on-safety-20

The MathWorks, I. (2022). *Fuzzy logic toolbox user's guide*. The MathWorks, Inc.

Thomas, K. W. (1992). Conflict and conflict management: Reflections and update. *Journal of Organizational Behavior*, *13*(3), 265–274.

Udugama, I. A., Gargalo, C. L., Yamashita, Y., Taube, M. A., Palazoglu, A., Young, B. R., Gernaey, K. V., Kulahci, M., & Bayer, C. (2020). The role of big data in industrial (bio) chemical process operations. *Industrial and Engineering Chemistry Research*, *59*(34), 15283–15297. https://doi.org/10.1021/acs.iecr.0c01872

Vaccari, M., Bacci Di Capaci, R., Brunazzi, E., Tognotti, L., Pierno, P., Vagheggi, R., & Pannocchia, G. (2021). Optimally managing chemical plant operations: An example oriented by Industry 4.0 paradigms. *Industrial and Engineering Chemistry Research*, *60*(21), 7853–7867. https://doi.org/10.1021/acs.iecr.1c00209

Vagia, M., Transeth, A. A., & Fjerdingen, S. A. (2016). A literature review on the levels of automation during the years. What are the different taxonomies that have been proposed? *Applied Ergonomics*, *53*, 190–202. https://doi.org/10.1016/j.apergo.2015.09.013

Vanderhaegen, F. (2021). Heuristic-based method for conflict discovery of shared control between humans and autonomous systems - A driving automation case study. *Robotics and Autonomous Systems*, *146*, 103867. https://doi.org/10.1016/j.robot.2021.103867

Vicentini, F. (2020). Terminology in safety of collaborative robotics. *Robotics and Computer-Integrated Manufacturing*, *63*(November 2019), 101921. https://doi.org/10.1016/j.rcim.2019.101921

Vinciarelli, A., Esposito, A., André, E., Bonin, F., Chetouani, M., Cohn, J. F., Cristani, M., Fuhrmann, F., Gilmartin, E., Hammal, Z., Heylen, D., Kaiser, R., Koutsombogera, M., Potamianos, A., Renals, S., Riccardi, G., & Salah, A. A. (2015). Open challenges in modelling, analysis and synthesis of human behaviour in human–human and human–machine interactions. *Cognitive*

Computation, *7*(4), 397–413. https://doi.org/10.1007/s12559-015-9326-z

Visser, E. J. de, J., E., Pak, R., & Shaw, T. H. (2018). From 'automation' to 'autonomy': the importance of trust repair in human–machine interaction. *Ergonomics*, *61*(10), 1409–1427. https://doi.org/10.1080/00140139.2018.1457725

Vogel, J., Haddadin, S., Jarosiewicz, B., Simeral, J. D., Bacher, D., Hochberg, L. R., Donoghue, J. P., & Van Der Smagt, P. (2015). An assistive decision-and-control architecture for force-sensitive hand-arm systems driven by human-machine interfaces. *International Journal of Robotics Research*, *34*(6), 763–780. https://doi.org/10.1177/0278364914561535

Wall, J. A., & Callister, R. R. (1995). Conflict and its management. *Journal of Management*, *21*(3), 515–558. https://doi.org/10.4324/9781003006039-2

Walliser, J. C., de Visser, E. J., Wiese, E., & Shaw, T. H. (2019). Team structure and team building improve human–machine teaming with autonomous agents. *Journal of Cognitive Engineering and Decision Making*, *13*(4), 258–278. https://doi.org/10.1177/1555343419867563

Wanasinghe, T. R., Trinh, T., Nguyen, T., Gosine, R. G., James, L. A., & Warrian, P. J. (2021). Human centric digital transformation and operator 4.0 for the oil and gas industry. *IEEE Access*, *9*, 113270–113291. https://doi.org/10.1109/ACCESS.2021.3103680

Wang, M., Wang, T., Luo, Y., He, K., Pan, L., Li, Z., Cui, Z., Liu, Z., Tu, J., & Chen,

X. (2021). Fusing stretchable sensing technology with machine learning for human–machine interfaces. *Advanced Functional Materials*, *31*(39), 1–13. https://doi.org/10.1002/adfm.202008807

Wen, H., Amin, M. T., Khan, F., Ahmed, S., Imtiaz, S., & Pistikopoulos, S. (2022). A methodology to assess human-automated system conflict from safety perspective. *Computers & Chemical Engineering*, *165*, 107939. https://doi.org/10.1016/j.compchemeng.2022.107939

Westera, W., Prada, R., Mascarenhas, S., Santos, P. A., Dias, J., Guimarães, M., Georgiadis, K., Nyamsuren, E., Bahreini, K., Yumak, Z., Christyowidiasmoro, C., Dascalu, M., Gutu-Robu, G., & Ruseti, S. (2020). Artificial intelligence moving serious gaming: Presenting reusable game AI components. *Education and Information Technologies*, *25*(1), 351–380. https://doi.org/10.1007/s10639-019-09968-2

Weyer, J., Fink, R. D., & Adelt, F. (2015). Human-machine cooperation in smart cars. An empirical investigation of the loss-of-control thesis. *Safety Science*, *72*, 199–208. https://doi.org/10.1016/j.ssci.2014.09.004

Woods, D. D., & Sarter, N. B. (1998). *Learning from automation surprises and "going sour" accidents: Progress on human-centered automation* (Issue January 2000).

Wu, H. N., Zhang, X. M., & Li, R. G. (2021). Synthesis with guaranteed cost and less human intervention for human-in-the-loop control systems. *IEEE Transactions*

on Cybernetics, 1–11. https://doi.org/10.1109/TCYB.2020.3041033

Wu, Z., Albalawi, F., Zhang, J., Zhang, Z., Durand, H., & Christofides, P. D. (2018). Detecting and handling cyber-attacks in model predictive control of chemical processes. *Mathematics*, *6*(10), 1–22. https://doi.org/10.3390/math6100173

Ylmaz, E. N., Ciylan, B., Gönen, S., Sindiren, E., & Karacayilmaz, G. (2018). Cyber security in industrial control systems: Analysis of DoS attacks against PLCs and the insider effect. *Proceedings - 2018 6th International Istanbul Smart Grids and Cities Congress and Fair, ICSG 2018*, *April*, 81–85. https://doi.org/10.1109/SGCF.2018.8408947

Yung, S. K., & Clarke, D. W. (1989). Local sensor validation. *Measurement and Control*, *22*(5), 132–140. https://doi.org/10.1177/002029408902200502

Zanchettin, A. M., Ceriani, N. M., Rocco, P., Ding, H., & Matthias, B. (2016). Safety in human-robot collaborative manufacturing environments: Metrics and control. *IEEE Transactions on Automation Science and Engineering*, *13*(2), 882–893. https://doi.org/10.1109/TASE.2015.2412256

Zero, L., Bersani, C., Paolucci, M., & Sacile, R. (2019). Two new approaches for the bi-objective shortest path with a fuzzy objective applied to HAZMAT transportation. *Journal of Hazardous Materials*, *375*(January), 96–106. https://doi.org/10.1016/j.jhazmat.2019.02.101

Zhang, Q., Langari, R., Tseng, H. E., Filev, D., Szwabowski, S., & Coskun, S. (2019). A game theoretic model predictive controller with aggressiveness estimation for

mandatory lane change. *IEEE Transactions on Intelligent Vehicles*, *5*(1), 75–89.

Zhao, J., & Zhang, X. (2020). Inverse tangent functional nonlinear feedback control and its application to water tank level control. *Processes*, *8*(3). https://doi.org/10.3390/pr8030347

Zhou, G. (2021). Human-Machine Cooperation and Path Planning for Complex Road Conditions. *Scientific Programming*, *2021*. https://doi.org/10.1155/2021/7262281