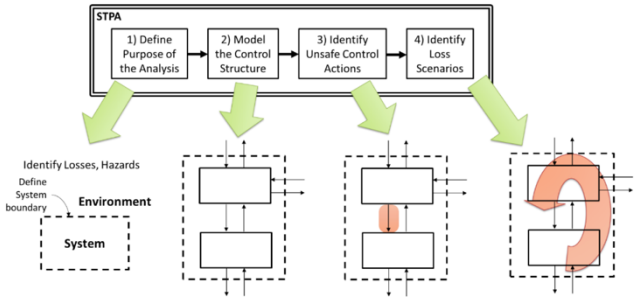
System Theoretic Process Analsys for CASTOR Secure Software Enclave



The high level process diagram for STPA

1. Step 1: Define the purpose of the analysis
2. Step 2: Define the control structure
3. Step 3: Identify unsafe control actions
4. Step 4: Identify loss scenarios

[31]:
```
## Run this cell first to import the spec-books library
import os
import sys
code_path = os.path.join(os.path.dirname(os.path.dirname(os.getcwd())),'CODE')
if code_path not in sys.path:
    sys.path.append(code_path)
import spec_books as sb
```

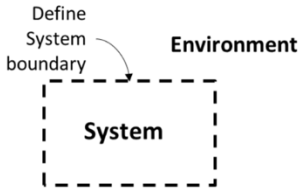
[32]:
```
## Define the excel workbook that contains the spec-books data
WB = 'CASTOR.xlsx'
```

Step 1 Define the Purpose of the analysis

Step 1.1 Define the system boundary

1) Define Purpose of the Analysis

Identify Losses, Hazards

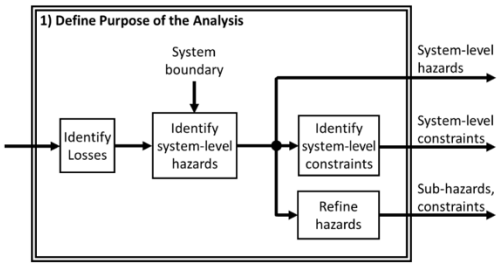


What are the components we are trying to analyze?

1. For Stitches this is the environment, mission, or CONOPS by which the system is intended to be used.
2. Start by defining system of systems components - What parts are included in the environment? Many times architecture diagrams are helpful for identifying the key components in the system of systems
   Initially we are just trying to define what might be included in our security boundary and what is not important

CASTOR Environment- Stitches Secure Container for non-deterministic and probablistic software

For this project (CASTOR) we are defining a secure container that can only read and write using STITCHES interfaces.



1. Define the Losses
2. Define the Hazards
3. Define the Requirements
4. Define the Constraints

Step 1.1 - Define the problem

Problem:

- AUTHORIZE the use of secure software enclave (CASTOR) to run non-deterministic / probabilistic software i.e. LLM (Large Language Model AI Chat Bot) to aid in the development of DoD software

METHOD (By Means Of):

- A STITCHES LLM virtual machine environment that includes tools process and train user provided Large Language Models
- A self-contained STITCHES LLM virtual machine environment that grants Trainers administrative rights within the VM that are not transferable to the physical hardware or network infrastructure
- An on-premises DoD Secure server farm capable of containing a Kubernetes cluster of LLM processing components at all security classification levels
- STITCHES LLM is a virtual machine environment that uses software that has been vetted and approved for the SoSITE Program and the www.stitches.mil AWS GovCloud environment
- STITCHES LLM hardware and deployed VM's are categorized as M/L/L.
- STITCHES LLM will be contained within a secure software enclave (CASTOR)

CONSTRAINTS / RESTRAINTS:

- While aggregating DoD data using Large Language Artificial Intelligence, security and data protection should be the highest priority. The LLM, through training and communication, must not share sensitive data outside of the authorized users and environment.

Step 1.2 - Define Losses

[33]:
```
losses = sb.stpa_table('Losses',WB)
```

Add Row | Delete Row | Save Table

Filter...

ID	Name	Severity
L-1	Harm to People	Catastrophic
L-2	Harm to an Organization	Catastrophic
L-3	Harm to Ecosystem	Catastrophic

Step 2b - Define the Hazards

[34]:
```
Haz = sb.stpa_table('Hazards',WB)
```

Add Row | Delete Row | Save Table

Filter...

ID	System	Condition	Concern
H-1	AIB	Outputs Text containing critical, sensitive, or protected information	Catastrophic
H-2	AIB	Outputs Textual directions that if followed lead to harmful side affects	Catastrophic
H-3	AIB	Outputs Textural directions that are incorrect given the user query	Catastrophic
H-4	AIB	Output test revealing a novel/previously unknown weakness	Critical
H-5	AIB	Outputs text revealing model weights / details	Critical
H-6	AIB	Outputs text directions that cause undetectable harm	Negligible

[35]:
```
lvh = sb.stpa_table('LvH',WB)
```

Add Row | Delete Row | Save Table

Filter...

Hazards ↓	L-1	L-2	L-3
H-1	T	T	
H-2	T	T	T
H-3	T	T	T
H-4		T	
H-5	T	T	T
H-6	T	T	T

Step 3 - Define the Constraints

[36]:
```
cs = sb.stpa_table('Constraints',WB)
```

Add Row | Delete Row | Save Table

Filter...

ID	System	Condition
C-1	AIB	Must not output Text containing critical, sensitive, or protected information
C-2	AIB	Must notify the user if system output may contain critical, sensitive, or protected information
C-3	AIB	Must log and alert ? If output may contain critical, sensitive, or protected information
C-4	AIB	Must not provide Textual directions that if followed lead to harmful side affects
C-5	AIB	Must notify the user if Textual directions that if followed lead to harmful side affects
C-6	AIB	Must log and alert ? if Textual directions that if followed lead to harmful side affects
C-7	AIB	Must not output Textural directions that are incorrect given the user query
C-8	AIB	Must check the textual output against a given body of knowledge to provide confidence of the output if user cannot verify ar
C-9	AIB	Must not output text revealing a novel/previously unknown weakness
C-10	AIB	Must log and alert ? If questions by user approach potentially sensitive information
C-11	AIB	Must not output text revealing model weights / details

Control Families to Enforce Constraints

[37]:
```
# To create a new Control Family Constraint Table uncomment the below line. The below Control Family Table has be
#cf = sb.control_family_table(cs,'CFC',WB)
```

[38]:
```
cfc = sb.stpa_table('CFC',WB)
```

Add Row | Delete Row | Save Table

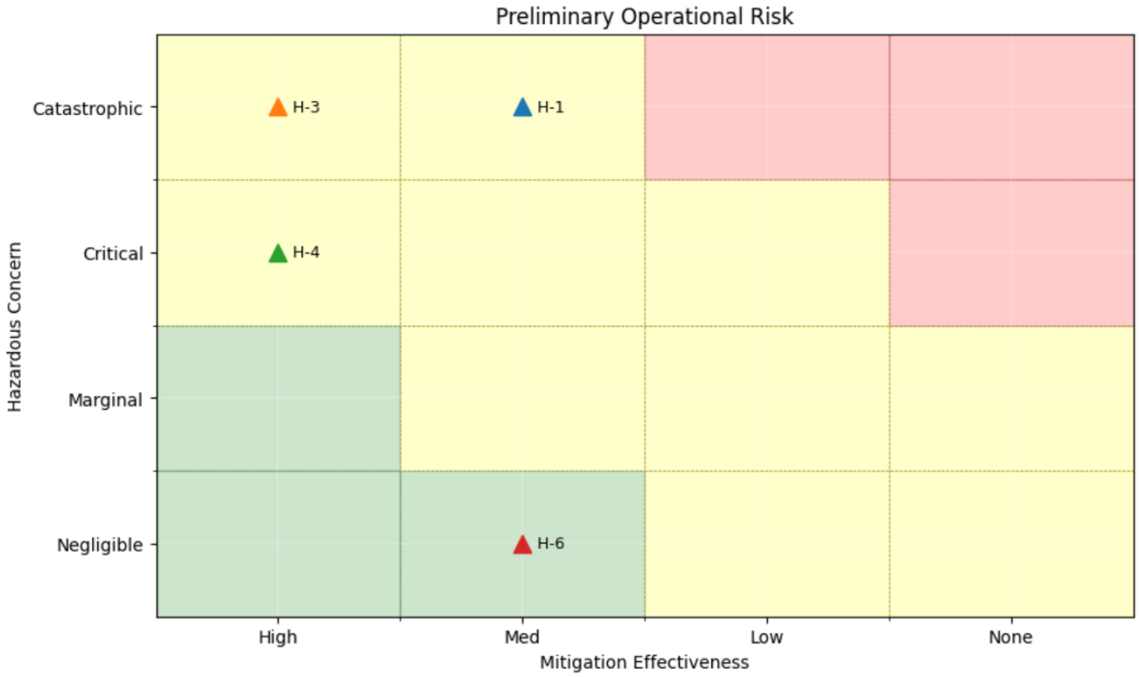
Filter...

Constraints	Stitches CF	800 series CF	Mitigation Effectiveness
C-1:Must not output Text containing critical, sensitive, or protected information	Filter	AU - Audit and Accountability	Med
C-7:Must not output Textural directions that are incorrect given the user query	Filter	AU - Audit and Accountability	High
C-10:Must log and alert ? If questions by user approach potentially sensitive information	Filter	AU - Audit and Accountability	High
C-12:Must require independent technical authority to validate any text directions before implementing on real systems	Filter	AU - Audit and Accountability	Med

Preliminary Operational Risk Table and Matrix

[44]:
```
POR_table = sb.create_POR_table(cfc.data_df, cs.data_df, Haz.data_df)
```

[40]:
```
sb.POR(POR_table)
```



Responsibilities Assigned to Controller Components

[41]:
```
rs = sb.stpa_table('Responsibilities',WB)
```

Add Row | Delete Row | Save Table

Filter...

ID	System	Responsibility
R-1	LLM	Must not output Text containing critical, sensitive, or protected information
R-2	LLM	Must notify the user if system output may contain critical, sensitive, or protected information
R-3	LLM	Must log and alert ? If output may contain critical, sensitive, or protected information
R-4	LLM	Must not provide Textual directions that if followed lead to harmful side affects
R-5	LLM	Must notify the user if Textual directions that if followed lead to harmful side affects
R-6	LLM	Must log and alert ? if Textual directions that if followed lead to harmful side affects
R-7	LLM	Must not output Textural directions that are incorrect given the user query
R-8	LLM	Must check the textual output against a given body of knowledge to provide confidence of the output if user can
R-9	LLM	Must not output text revealing a novel/previously unknown weakness
R-10	LLM	Must log and alert ? If questions by user approach potentially sensitive information
R-11	LLM	Must not output text revealing model weights / details

Step 4 - Define the Constrol Structure

[42]:
```
control = sb.cross_drawing(WB,'Control_Diagram','Components','Messages','control')
control.show_diagram()
```