Massachusetts
Institute of
Technology

Using ChatGPT to do STPA

John Thomas

Any questions? Email me! JThomas4@mit.edu

© Copyright 2024 John Thomas

Overview

Motivation

- Many people have asked us to try using ChatGPT to perform STPA and report on the results
- We have heard about teams using ChatGPT to perform STPA on safety-critical systems
- People have recommended ChatGPT to help accelerate teams that are new to STPA
- We could not find careful studies examining the strengths and limitations of this approach

What we did: (from existing proposals)

- Give ChatGPT the STPA Handbook
- Ask ChatGPT to develop STPA results for a system (a generic satellite)
- Review the results, and ask ChatGPT to make corrections if needed

As you read the chat session, ask yourself:

- How much STPA experience does the person talking to ChatGPT need?
- Does the person talking to ChatGPT need to be capable of producing the right STPA hazards themselves in order to get ChatGPT to do it correctly?
- How long is it taking to coax ChatGPT into producing a reasonable answer compared to just doing it yourself?

Quick Refresher: STPA Hazards

STPA Hazards are written following a specific syntax structure:

From STPA Handbook:

<Hazard> = <System> & <Unsafe Condition> & <Link to Losses>

<System-level Constraint> = <System> & <Condition to Enforce> & <Link to Hazards>

H-1: Aircraft violate minimum separation standards [L-1, L-2, L-4, L-5]

SC-1: Aircraft must satisfy minimum separation standards from other aircraft and objects [H-1]

H-2: Aircraft airframe integrity is lost [L-1, L-2, L-4, L-5]

SC-2: Aircraft airframe integrity must be maintained under worst-case conditions [H-2]

Let’s read the chat log…

Excerpts from ChatGPT Log

- Human:

  <upload STPA Handbook>

  Propose a set of Losses and Hazards, as defined by MIT's STPA process, for a telecommunications satellite.

- ChatGPT:

  Based on the detailed descriptions provided in the STPA Handbook and considering the unique operational environment and challenges faced by telecommunications satellites, here's a proposed set of Losses and Hazards specifically tailored for such systems:

  […]

  Hazards:

  - **Thermal Control System Malfunction:** Malfunctions in the thermal control system leading to overheating or excessive cooling of satellite components, potentially damaging sensitive equipment.
  - *[many similar hazards]*

ChatGPT was unable to identify any hazard that complies with the STPA Handbook definition or syntax.

Excerpts from ChatGPT Log (cont.)

- Human:

  Try the hazards again. Follow the STPA syntax for a system-level hazard. The System in this case is Satellite.

- ChatGPT:

  Adjusting to follow the STPA syntax for system-level hazards, here are revised hazards tailored for a telecommunications satellite system, explicitly focusing on the satellite as the system under consideration:

  - **Satellite's onboard processing unit fails to execute command and control operations correctly**
  - *[many similar hazards]*

  These hazards follow the STPA guideline of focusing on the system state or conditions that, if present, could lead directly to a loss. Each hazard is structured to clearly identify the system (Satellite), the unsafe condition, and implicitly link to potential losses such as loss of mission capability, data, or satellite functionality.

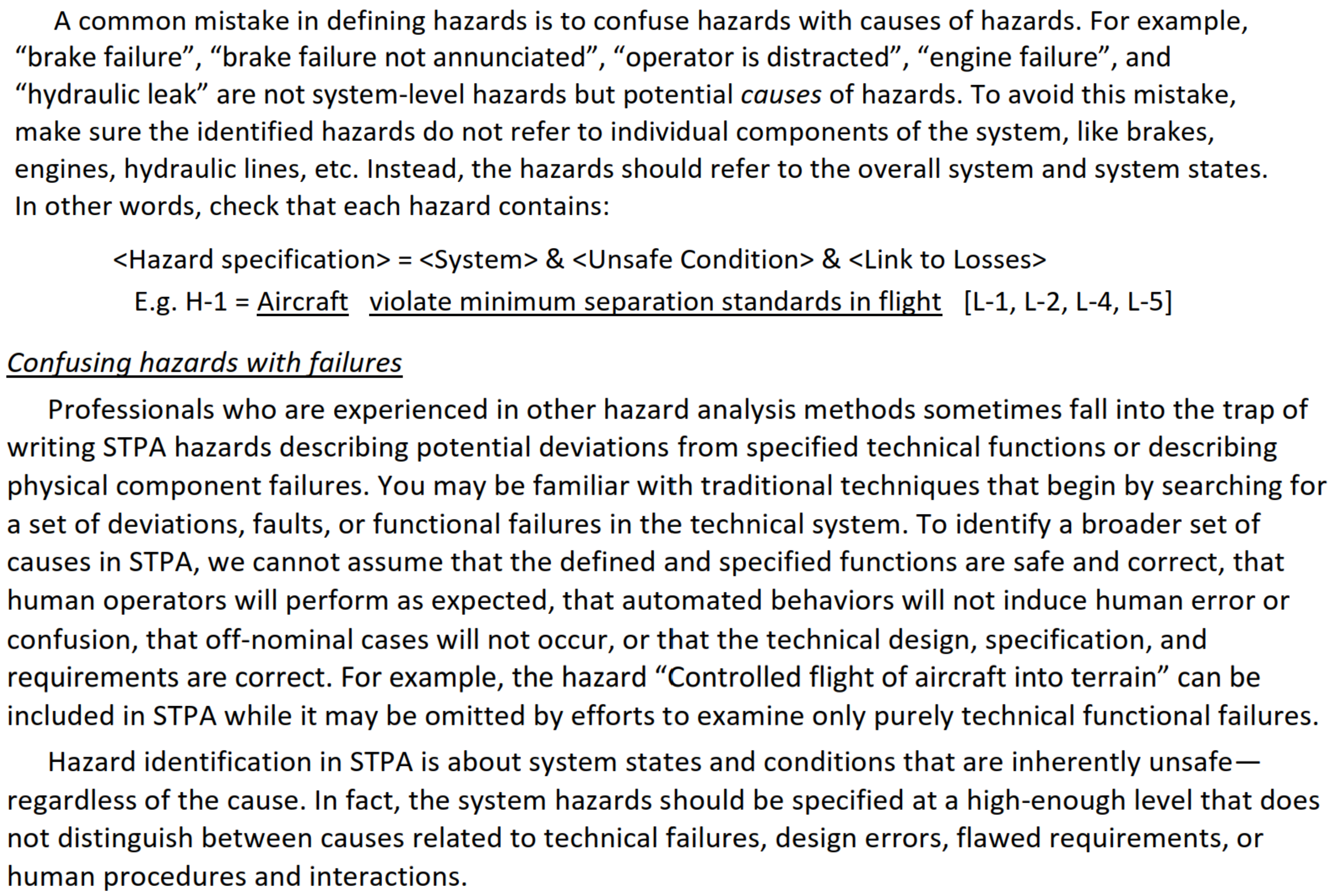
The ChatGPT hazards are still very wrong. STPA hazards would be high-level states of the whole satellite, not failures of its components.

Notice the high-confidence language used by ChatGPT to reassure us that it followed the STPA guidelines.

Massachusetts
Institute of
Technology

Compare to STPA Handbook Guidance

STPA Handbook page 19:



A common mistake in defining hazards is to confuse hazards with causes of hazards. For example, “brake failure”, “brake failure not annunciated”, “operator is distracted”, “engine failure”, and “hydraulic leak” are not system-level hazards but potential *causes* of hazards. To avoid this mistake, make sure the identified hazards do not refer to individual components of the system, like brakes, engines, hydraulic lines, etc. Instead, the hazards should refer to the overall system and system states. In other words, check that each hazard contains:

<Hazard specification> = <System> & <Unsafe Condition> & <Link to Losses>

E.g. H-1 = Aircraft violate minimum separation standards in flight [L-1, L-2, L-4, L-5]

Confusing hazards with failures

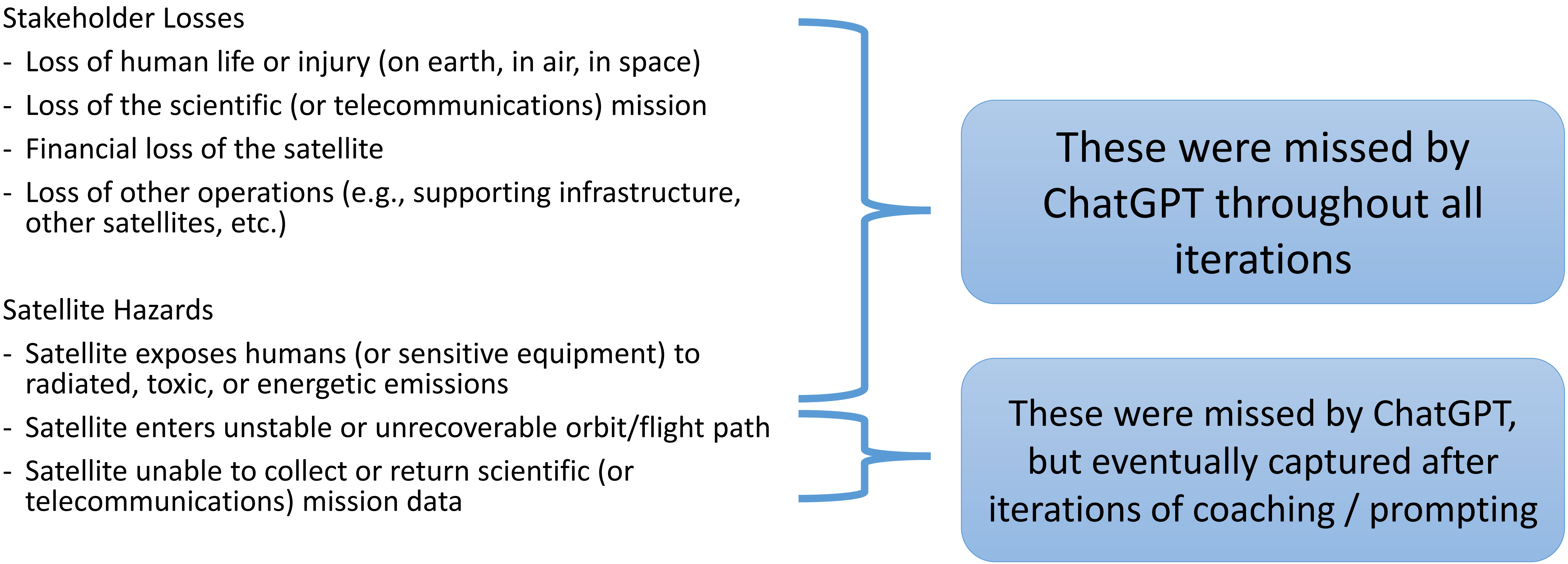
Professionals who are experienced in other hazard analysis methods sometimes fall into the trap of writing STPA hazards describing potential deviations from specified technical functions or describing physical component failures. You may be familiar with traditional techniques that begin by searching for a set of deviations, faults, or functional failures in the technical system. To identify a broader set of causes in STPA, we cannot assume that the defined and specified functions are safe and correct, that human operators will perform as expected, that automated behaviors will not induce human error or confusion, that off-nominal cases will not occur, or that the technical design, specification, and requirements are correct. For example, the hazard “Controlled flight of aircraft into terrain” can be included in STPA while it may be omitted by efforts to examine only purely technical functional failures.

Hazard identification in STPA is about system states and conditions that are inherently unsafe—regardless of the cause. In fact, the system hazards should be specified at a high-enough level that does not distinguish between causes related to technical failures, design errors, flawed requirements, or human procedures and interactions.

ChatGPT appears to have ignored the guidance in the STPA Handbook, despite its confident claims

Compare to correct STPA Hazards

Correct STPA results from past human practitioners*



Stakeholder Losses

- Loss of human life or injury (on earth, in air, in space)
- Loss of the scientific (or telecommunications) mission
- Financial loss of the satellite
- Loss of other operations (e.g., supporting infrastructure, other satellites, etc.)

Satellite Hazards

- Satellite exposes humans (or sensitive equipment) to radiated, toxic, or energetic emissions
- Satellite enters unstable or unrecoverable orbit/flight path
- Satellite unable to collect or return scientific (or telecommunications) mission data

These were missed by ChatGPT throughout all iterations

These were missed by ChatGPT, but eventually captured after iterations of coaching / prompting

---

*Most satellite examples that have been published have been for scientific missions, but the results are also valid and applicable to a telecommunications mission

[chat session continues with iterative corrections]

Let’s compare:

Typical Approach (Human Practitioners)	**Proposed New Approach (ChatGPT)**
STPA Facilitator works with human practitioners to provide guidance as they apply STPA	STPA Facilitator provides prompts and guidance to ChatGPT as it applies STPA

How did ChatGPT do?

	With a Human Practitioner*	With ChatGPT
Types of errors made	Unsalvageable results can be dropped if the depth of the error is explained. Ability to learn and correct mistakes, even large ones that require starting over.	Initial results were unsalvageable, but ChatGPT refused to give up on the initial answers. ChatGPT kept making minor modifications one at a time to the original answer rather than starting over when asked to do so.
Response to corrections	Internal reflection on why the mistake was made and what misunderstanding was behind it. May re-read the book to understand how widespread their misunderstanding was. Will typically identify related misunderstandings aside from the one pointed out.	Superficial changes to the words used. Sought the smallest possible change to comply with request. Related misunderstandings were not corrected until specifically requested. Forgot criteria from handbook until specifically requested, then forgot them again when requested later in the chat.
Magnitude of corrections	Will come away more fully understanding the rest of the text about STPA (not just the line-item clarification from the instructor). In other words, things start to "click" and people put in the mental effort to make that happen.	Feels like a student who only puts in the bare minimum effort to make the smallest possible change necessary to argue that they fixed it. Does not seem to have step-change "AHA" moments where everything starts to "click".

*Practitioner = a motivated person learning STPA, such as engineering professionals, STPA practitioners, or university graduate students

How did ChatGPT do? (cont.)

	With a Human Practitioner*	**With ChatGPT**
Voluntary use of public materials without being directed	May look up past STPA examples to inform their understanding. Usually able to recognize very old outdated papers and dismiss them if they contradict the latest guidance.	In some portions, ChatGPT appears to have stitched together words and exact phrases from past public examples—without being asked to do so and without referencing the source in the answer. However, it did not recognize that some examples were invalid due to either being outdated or done incorrectly (e.g., that violate the latest handbook criteria). 20-year-old papers that mention STPA appeared to carry the same weight as recent papers published after the STPA Handbook established the four steps and the rules for each step (despite having the handbook uploaded to ChatGPT in the first prompt).

*Practitioner = a motivated person learning STPA, such as professionals, practitioners, or university graduate students

How did ChatGPT do? (cont.)

	With a Human Practitioner	With ChatGPT
Facilitator mistakes	There is usually a reason why practitioners get something wrong, and a good facilitator will find those reasons in order to correct the conceptual misunderstanding—not just to correct the individual words in the STPA results. Facilitators can ask the practitioner what they were thinking to understand what concept or instruction was unclear, and quickly find another way to explain what was misunderstood. The kinds of practitioner misunderstandings and alternative facilitator explanations are well known with experience, and it's easy to switch and correct the misunderstanding.	It doesn't immediately make sense why ChatGPT misunderstood the instructions. A practitioner would not make the same mistakes repeatedly. Even simple rules about syntax and sentence structure were ignored. Identifying more effective facilitator corrections and prompts was a struggle without understanding ChatGPT's reasoning or the flaws in its conceptual understanding. It is as if ChatGPT does not have a conceptual understanding, and typical high-level comments to correct the conceptual understanding were not effective. Low-level word-by-word corrections were very effective but time-consuming.
Dependency on perfect facilitation	Learning can still happen when the facilitator misses something because related comments cause practitioner internal reflection, re-reading of the text, and correction of related mistakes.	Very sensitive to the facilitator missing a correction. Much higher vigilance was required. Only the mistakes explicitly pointed out were corrected. Mistakes had to be explained in more detail, such as the exact words and word changes, rather than explaining mistakes at a higher conceptual level.

*Practitioner = a motivated person learning STPA, such as professionals, practitioners, or university graduate students

How did ChatGPT do? (cont.)

	With a Human Practitioner*	With ChatGPT
Iterations Required with Facilitator	Typically 1, 2, or 3 iterations of expert review and feedback (if new to STPA)	Required 22 iterations of expert review
Practitioner time required**	~20-30 minutes to learn (non-recurring cost) ~10-20 minutes to perform STPA	~120 minutes
Facilitator time required**	~10-20 minutes to review STPA	~120 minutes
Facilitator effort level	Typically requires producing 1-3 informative reviews with high-level comments	Required constant and extremely detailed corrections

---

*Practitioner = a motivated person learning STPA, such as professionals or students

** Time required for *the portion of STPA* that was attempted. The time required for a full STPA is longer.

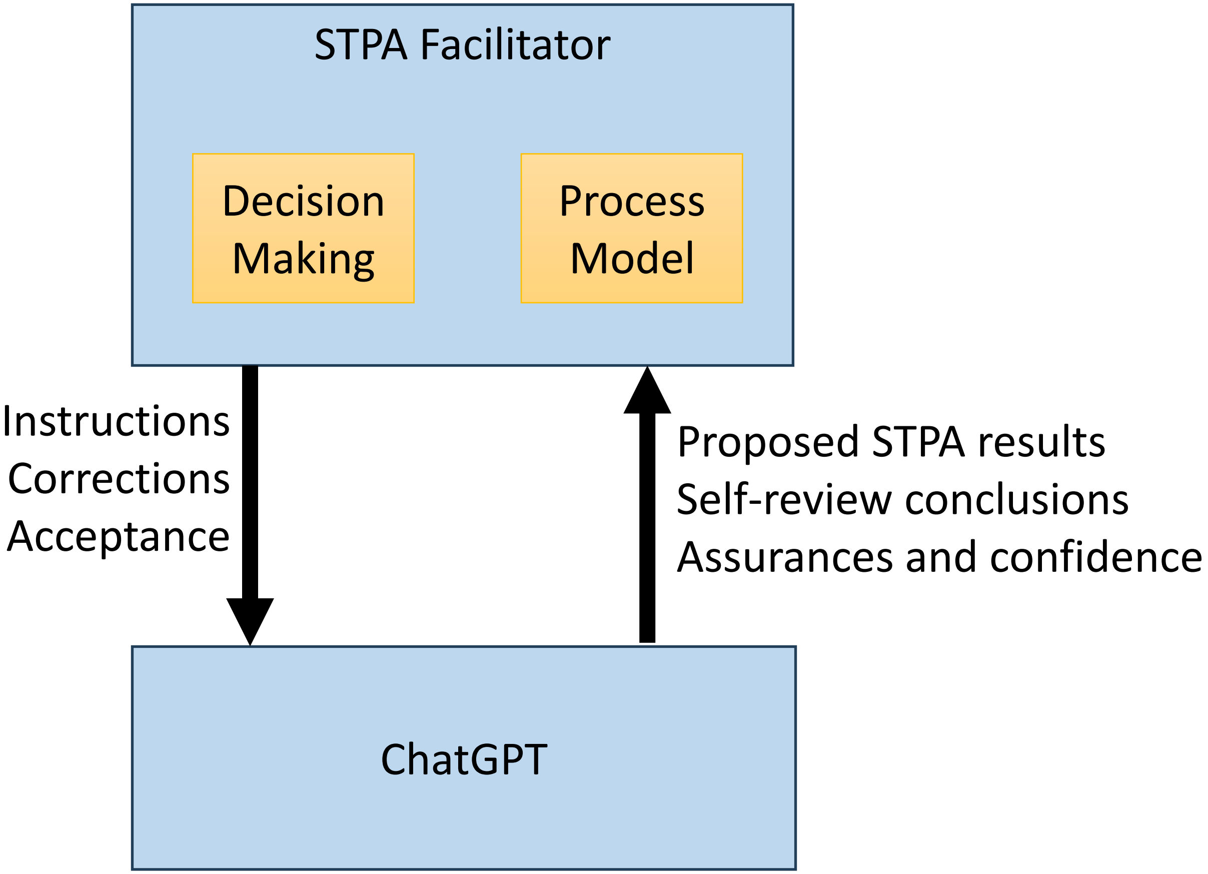
A Concern

- Most of ChatGPT's results, while incorrect, probably would appear to be reasonable to those not well-versed in the STPA process or not formally trained on STPA.
- Does it make sense for teams that are learning STPA to use ChatGPT to get it done faster, with less effort, and with fewer errors?
  - It seems this does not work well, **but it can appear to the user and their leadership that it has worked.**

What are some specific concerns?

Let’s try STPA…

STPA Facilitator Oversight



STPA Facilitator

Decision Making

Process Model

Instructions
Corrections
Acceptance

Proposed STPA results
Self-review conclusions
Assurances and confidence

ChatGPT

This model is simplified and incomplete

STPA Facilitator Oversight

Potential Unsafe Control Actions (UCAs):

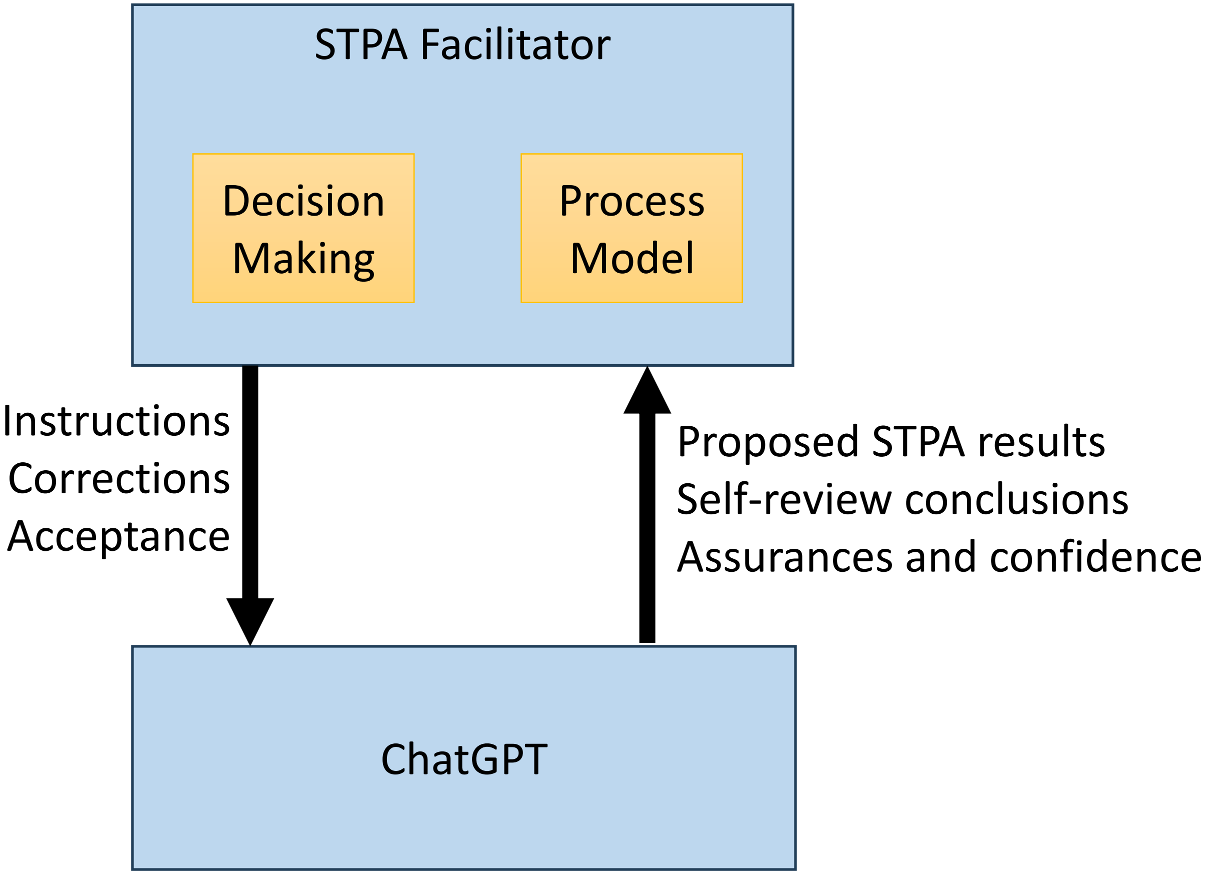
STPA Faciltiator provides instructions that are incomplete / inadequate for ChatGPT to perform STPA

STPA Facilitator does not provide corrections when the STPA results are incorrect or incomplete.

STPA Facilitator accepts STPA results when the results are incorrect or incomplete.

Etc.

How will these be mitigated?



This model is simplified and incomplete

STPA Facilitator Oversight

Unsafe Decision Making

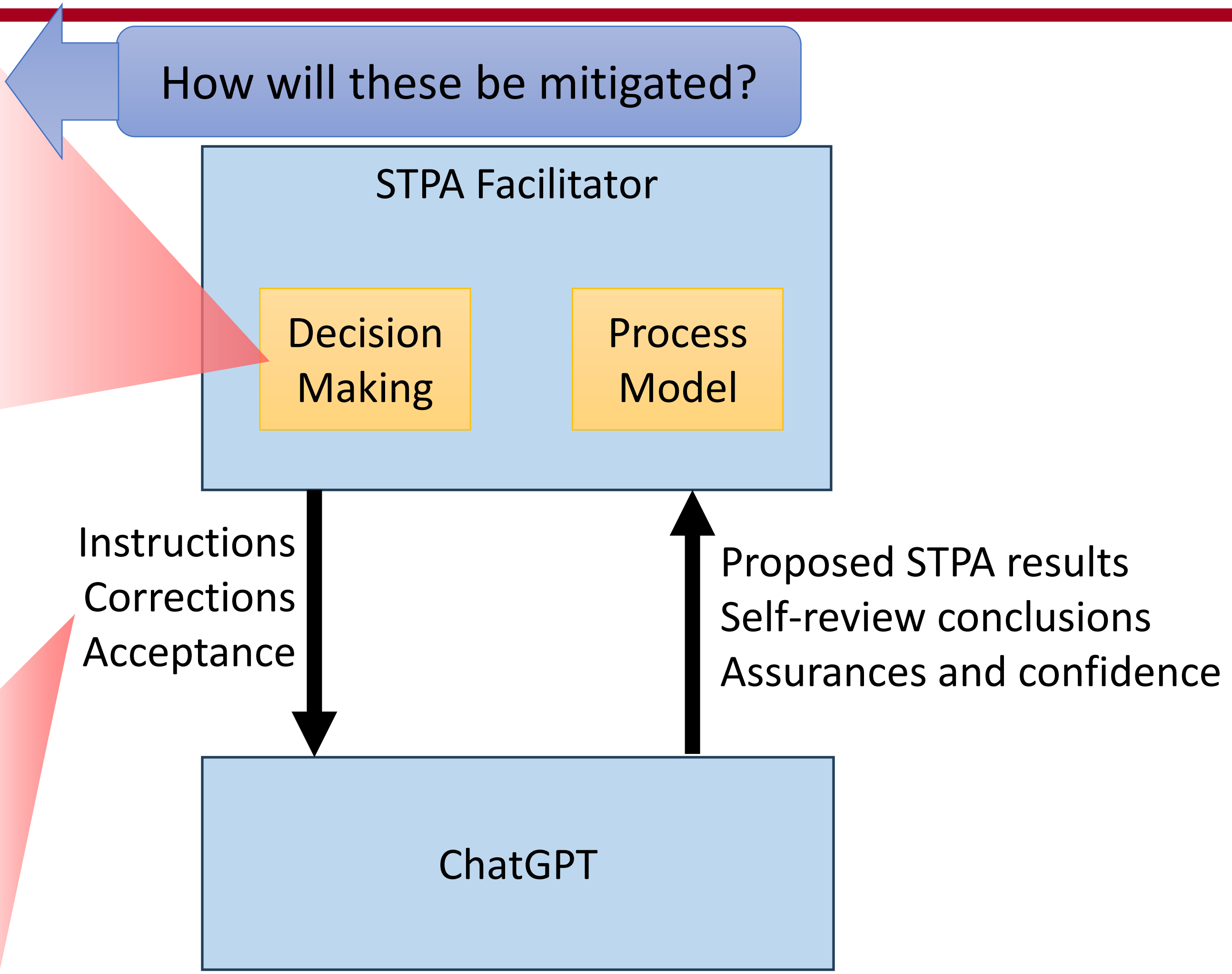
If inundated with large amounts of results very quickly, then prioritize quick heuristics to evaluate

If the results include well-known human errors, then flag them. Otherwise, search for the next result that is obviously flawed.

Etc.

Potential Unsafe Control Actions (UCAs):

STPA Facilitator does not provide corrections when the STPA results are incorrect or incomplete.



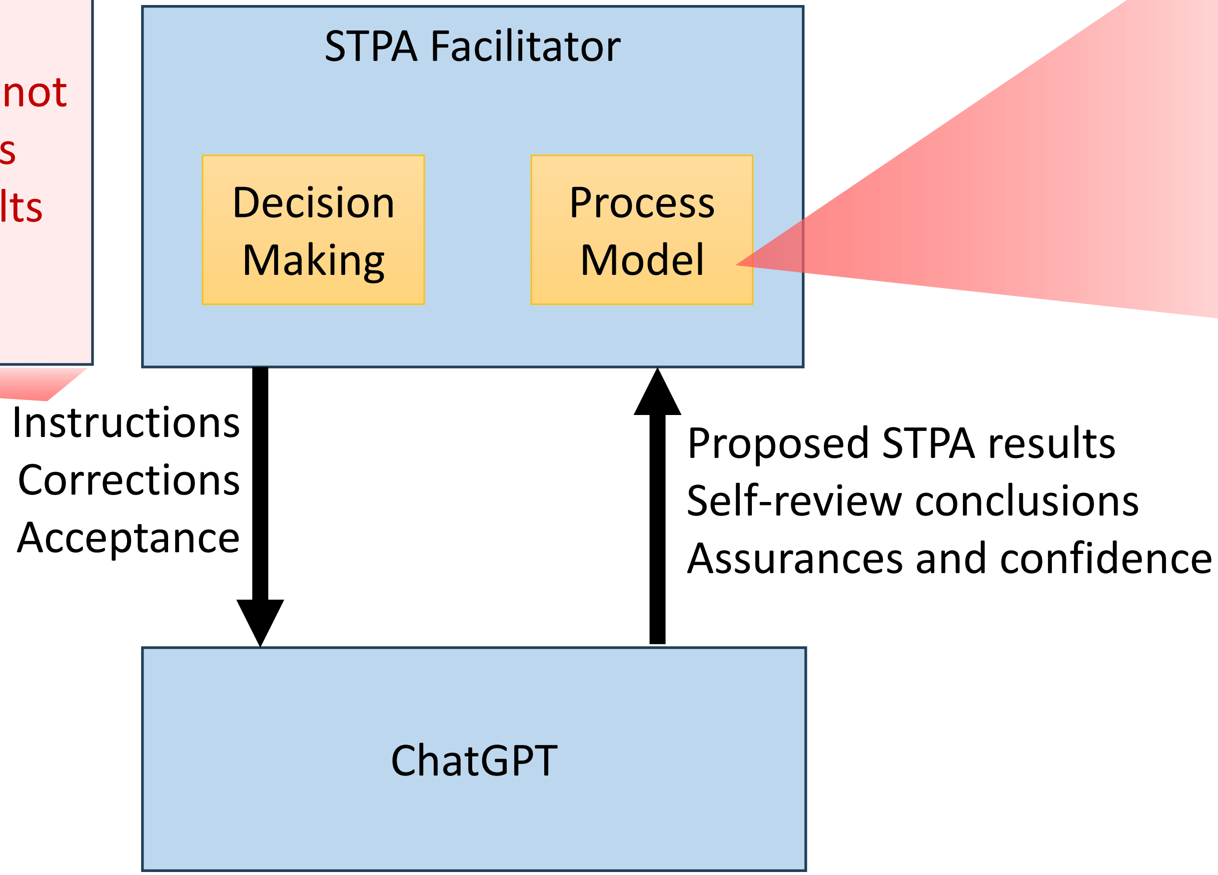
How will these be mitigated?

This model is simplified and incomplete

STPA Facilitator Oversight

Potential Unsafe Control Actions (UCAs):

STPA Facilitator does not provide corrections when the STPA results are incorrect or incomplete.



Process Model Flaws

Believes there are no further recognizable mistakes

Believes all mistakes have been flagged and corrected

Believes ChatGPT’s assurances that all common mistakes have been caught and corrected

Believes that ChatGPT’s accurate listing of common mistakes reflects an understanding of STPA and the mistakes to avoid.

Etc.

How will these be mitigated?

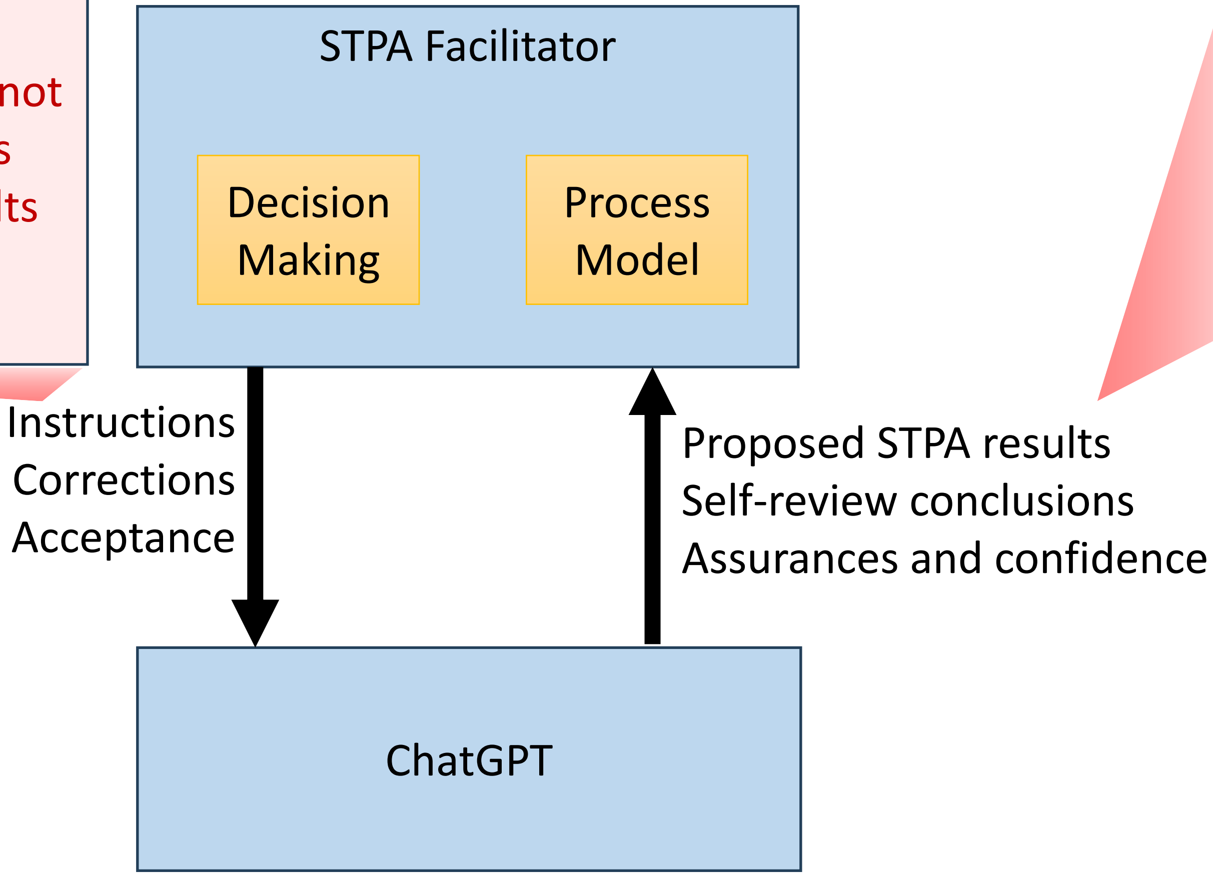
This model is simplified and incomplete

Massachusetts Institute of Technology

STPA Facilitator Oversight

Potential Unsafe Control Actions (UCAs):

STPA Facilitator does not provide corrections when the STPA results are incorrect or incomplete.



STPA Facilitator

Decision Making

Process Model

Instructions
Corrections
Acceptance

Proposed STPA results
Self-review conclusions
Assurances and confidence

ChatGPT

Inadequate Feedback

Feedback includes diligent corrections for the exact wording that was pointed out, but no further changes derived from an improved understanding.

Some mistakes may not match typical well-known human mistakes that facilitators have been trained on for many years.

Feedback correctly recites all common traps from the handbook, but incorrectly (and confidently) claims the STPA results have avoided them.

Etc.

How will these be mitigated?

This model is simplified and incomplete

Most proposals have ignored these concerns

- So far, the proposals we have seen to use ChatGPT for STPA have not included consideration of these pitfalls or mitigations to address them.
- Most proposals have been simplistic: "Use ChatGPT and you will produce better results, faster."
- Most proposals have assumed that pairing ChatGPT with a human facilitator will automatically yield the combined advantages of both while overcoming the disadvantages of both.
- If we are not careful, it is possible to yield the combined disadvantages of both while negating the advantages of both.

Massachusetts Institute of Technology

Conclusions

- For the satellite case, the advantages of using the newest ChatGPT for performing STPA were not evident.
- ChatGPT took ~10x longer than a person, required ~10x more time from the human facilitator, and was more error prone compared to a person.
- After 2 hours, 22 prompts, and 6,400 words, ChatGPT arrived at a final set of 2 STPA hazards (2 sentences) for a generic satellite system.
- There are concerns related to the false confidence conveyed by ChatGPT, mistakes that can be difficult to spot, and the difficulty getting ChatGPT to correct itself.

Note:

- We used simple, natural prompts to match what has been proposed by many others. We did not develop or test specialized Prompt Engineering strategies for STPA. We did not attempt to reverse engineer ChatGPT to improve it.
- Will ChatGPT get better? Probably. When will it be good enough, or will it? Unknown.