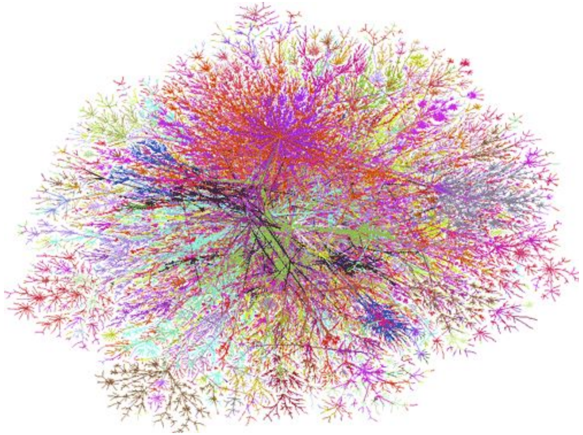
Computer Science Module 2:

Reliability & Liability

Prepared by Dr. Daniel Rosiak, Prof. Shannon E. French, and Beth Trecasa

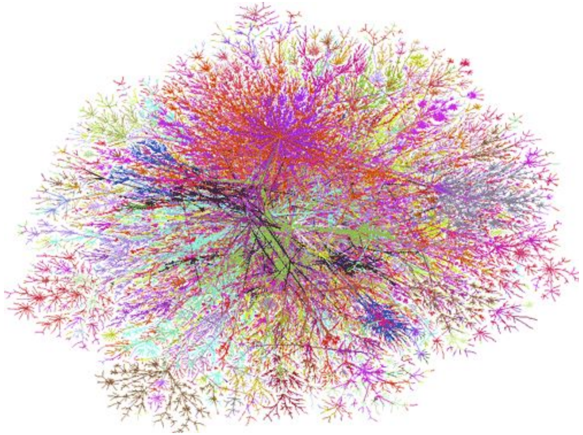
Introduction



It is difficult to overstate how dependent we are today on computers and computerized systems to facilitate so many of our daily activities. Such systems now govern most of modern communication, transportation, finance, retail, healthcare, military systems, and more.

When computerized systems work correctly, they can of course save us time, money, and enable the accomplishment of many other activities. But when they fail or something goes wrong, any of those benefits can quickly be overturned by rather disruptive or impactful harms. Failures of computer-integrated systems can ramify throughout many areas of society, resulting in lost time, lost money, social injustice, and even injury or loss of life.

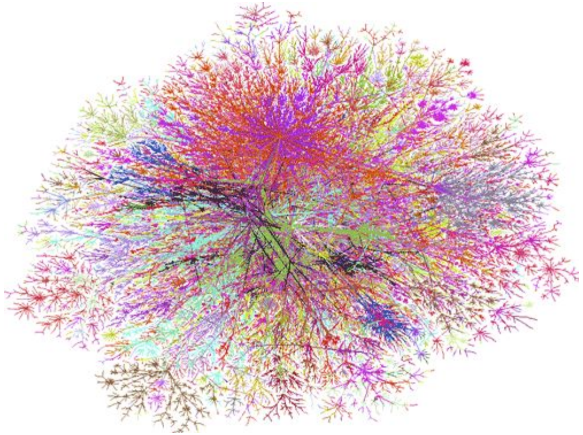
Introduction



Computers and code nearly always form part of larger systems – like financial systems, health-care systems, transportation systems, etc. – and fundamentally it is the reliability of the *entire system* that is most important.

A system that is designed well is one that can tolerate the malfunction of any single component without failing or causing harm.

Introduction



These two paired modules are designed to familiarize you with:

- some of the ways in which computerized systems have proven to be unreliable,
- what we can do to make them more reliable,
- what the broader ethical stakes are, and
- general practices we can implement to encourage greater awareness and anticipation of the risks.

Specific Objectives

The main purpose of this **two-part Reliability & Liability module is** to invite aspiring computer scientists and others involved in the construction, design, upkeep, and testing of large systems or system components to increase their awareness of some of the broader ethical risks associated with the ways things can go wrong. In many areas, more and more functions are being ceded to algorithms to the detriment of human control, without full awareness of the consequences. This raises a host of concern for loss of predictability, fairness, and equitability.

Module 1

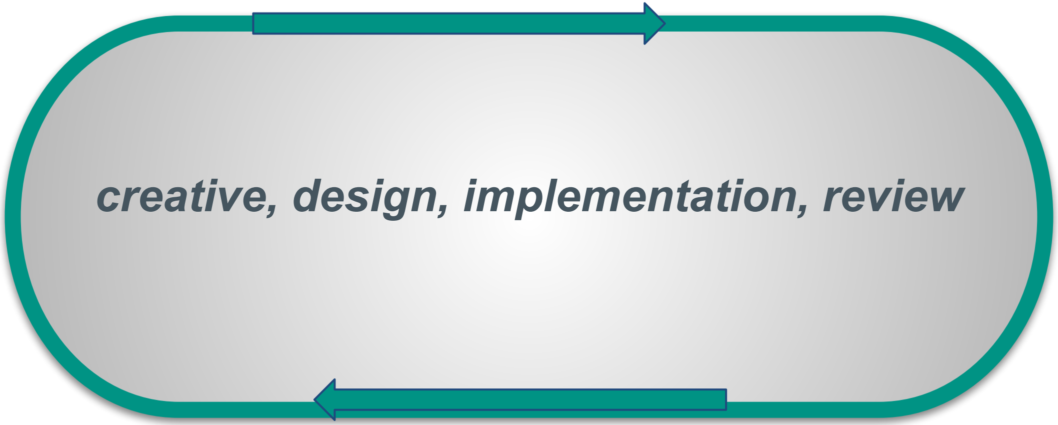
- A number of carefully-selected **examples** to illustrate **three main errors types**
- Important **lessons** drawn from each of those examples
- An **exercise** for students

Module 2

- General tools for better identifying and mitigating ethical risks
- An **exercise** for students
- Further **readings and resources**

- ☑ **Early**
- ☑ **Often**
- ☑ **Always**

Ethics in Computer Science & Engineering



creative, design, implementation, review

Identifying and Mitigating Ethical Risks **at Every Stage**

Topics Covered in Module 2

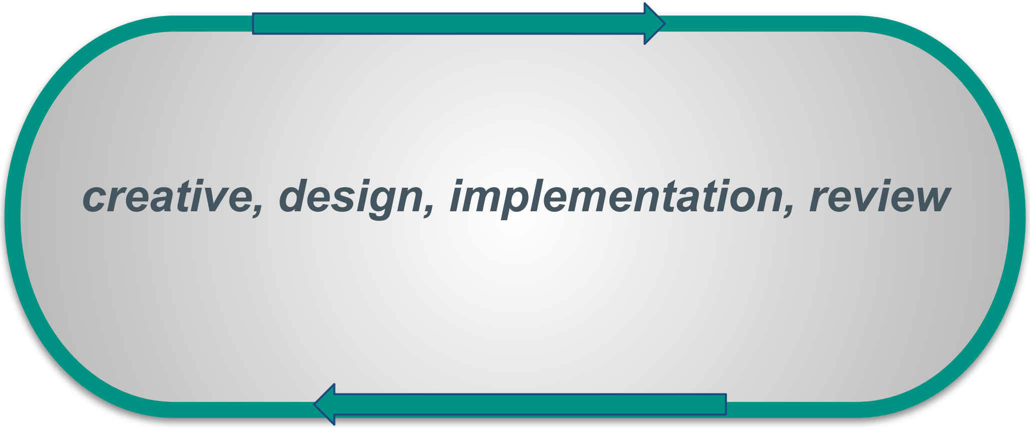
1. **Preliminary Considerations:** Ethics at Every Stage
2. **Routines:** Addressing Ethical Considerations in Computer Science and Engineering
3. **In Review:** Looking Back to Look Forward

1. Preliminary Considerations:
Ethics at Every Stage

Why a loop? Ethics work is never done.

In engineering and computer science, consideration of ethics should be concurrent with the:

- creative process,
- design process,
- implementation, and
- post-implementation review.



Consider this: Ethics at Every Stage

- ❑ Consider **the company you keep** and **who is in the room** where things happen.
- ❑ Consider **who will be affected** by what you are creating.
- ❑ Consider **how those affected will be affected**: what (if any) possible harms could come from what you are creating? What benefits are they likely to see?
- ❑ Consider **what you can build** into the existing design **to minimize or eliminate** those possible harms.
- ❑ Consider not just ways of avoiding harm, but also **ways of ensuring positive outcomes** and unlocking positive potentials.

Consider this: Ethics at Every Stage

Even when everything looks acceptable from your end, it is always important to:

❑ Consider what can happen **if "bad actors" abuse or misuse** what you have created – where "bad actors" include:

- people with bad or destructive intentions,
- people who dangerously misinterpret or find irrational uses for your creation,
- people who may weaponize what you built.

❑ Consider – in advance, instead of just responding to such abuses and misuses after the harm is done – how you can build things to **make such abuses and misuses less likely** from the start.

Consider this: Ethics at Every Stage

- ❑ Consider whether your **design has left room for corrections**, if and when something goes wrong with the system you have built.
- ❑ Consider when something goes wrong, if many more things will go wrong (a “snowball effect”).
- ❑ Consider **ethical feedback mechanisms** and iterate them. Ethics is not some external requirement to be checked off and then ignored until the next “inspection” or until the next disaster forces our hand.

Lesson Learned: *Ethics is about firmly grounding ourselves in the best possible outcomes in our field, and can help guide you throughout your ongoing endeavors to become the sort of computer scientists, engineers, designers you would like to be.*

Consider this: the company you keep

- ❑ **Remember** you are not the first person to attempt to build systems that are useful AND ethical. However, you may sometimes be the only voice for ethical considerations in the room at a given moment.
- ❑ **Take steps to ensure** you are never working in a bubble (especially with respect to any ethically-relevant components of your work). You are part of a community that can help you avoid making errors – especially those that could cause real harm. You are also part of a community that can engineer ethical solutions – especially those that could have real impact and help address old and new problems. Don't stay in a bubble - step out of it, and let others inside.

Consider this: the room where it happens

❑ **Remember to** learn from those who have come before, by remaining engaged, carefully considering relevant cases, and studying any “lessons learned.” Read case studies and continually seek out current examples to find out what has and has not worked well in the past and how the lessons might apply to your current work.

❑ **Take steps to ensure** you are always vigilant about who is a part of the conversation.

- Are there non-engineers in the room?
- Is there diversity in other ways?

Lesson Learned: *A diversity of perspectives will identify more potential positives and potential negatives*. *No one person or group can think of every possible impact.*

Consider this: why you are doing what you are doing

- ❑ **Remember** while focusing on *why* you are doing or creating something, also be sure to think carefully, in advance and often, about what you are willing (and may have) to sacrifice to do it right. For example, you may need to move a deadline to make a better product, and you should always put people before profit.

Lesson Learned: *Do this regularly in order to be prepared to stick up for what is right, when needed. Even good people, with good intentions, can go astray – you can steel yourself against making unethical compromises by regularly reflecting on why you do what you do.*

Consider this: ways of encouraging positive outcomes

- ❑ **Remember to** strive to do no harm, but also to continually do better. Being ethical is not only about *not doing things* that cause harm. It is also about *unlocking positive potential* and *achieving desirable ends/goals*.

- ❑ **Take steps to ensure** that when you create a new system, you are never limiting yourself to merely complying with existing rules and regulations or best practices. Think beyond limiting harm and look for innovative ways to do real good and help the people who will interact with what you have built to be the best people they can be.

Lesson Learned: *It is not enough to try to avoid introducing bias in a system (i.e., to avoid using a data set lacking sufficient diversity); the mark of a good system would be one that actively guided users towards desirable ends, such as fairness and inclusion.*

Ethics Involves Both Setting Limits AND Unlocking Potential:

Following Rules and Exercising Restraint

- Focus is on controlling human behavior so that humans can live together in successful social units
- Tends to presume a somewhat negative view of human nature – without the constraints of ethical guidelines, we will do very bad things to one another (and even to ourselves)

Seeking and Living the Good Life

- Focus is on developing character and habits that help humans reach their positive potential and flourish (both individually and collectively)
- Compatible with the idea that human nature, while flawed, can be improved – and may even be perfectible – that virtue is habit-forming

Ethics at Every Stage: Questions to Consider

□	Who is part of the discussion?
□	Who will or might be affected?
□	What additional information/facts might be required to make an assessment of the relevant effects?
□	What concrete steps can you take in order to gain access the information/perspectives needed?
□	What moral values are potentially conflicting with each other in this situation? Is there any way to reconcile them?
□	If conflict is unavoidable, are there ways to respect all relevant interests/values? If not, is there an argument that certain values ought to be prioritized over others?

Ethics at Every Stage: Questions to Consider

□ What are the possible harms?

□ Are there subtle risks you might have missed?

□ Are there bad actors who might misuse or abuse your creation?

□ Are there similar cases out there from which we can learn?

□ How can we reduce possible harms?

□ How can we reduce potential abuses?

□ How can we fix it and improve in the future?

2. Routines: Addressing Ethical Considerations in Computer Science and Engineering

Ethical risks

- In technology fields, our technical and design decisions can and often do have immense impact on a wide range of matters of moral import.
- In the history of technology design and engineering, a large number of mostly avoidable harms and disasters have arguably arisen from failing to adequately **identify the foreseeable relevant ethical risks** in a particular product or creation.
- ***Ethical risks*** can be broadly defined to include any choices that may cause substantial harm to persons or other entities or systems with a moral status (such as animals, the environment, democratic institutions), or that may lead to considerable controversy or dispute for other reasons.

Routines for Ethics?

- There exist a variety of theories/step-by-step approaches/principled systems for assessing the rightness/wrongness of something – and so, for helping us assess what might or might not constitute an ethical risk.

- The work developing these approaches, and the struggle between the various competing and overlapping frameworks – including the likes of Utilitarianism, Deontology, and Virtue Ethics – is the subject of general ethics.

- Ethical reasoning is a skill you can develop and hone, so that you get better at recognizing the ethical dimensions of a project and how to address them.

- Ethics courses (see **Further Readings and Resources**) are a great place for starting to explore these, should you be interested in learning more (which we hope you are)!

Ethical "Risk Sweeping"

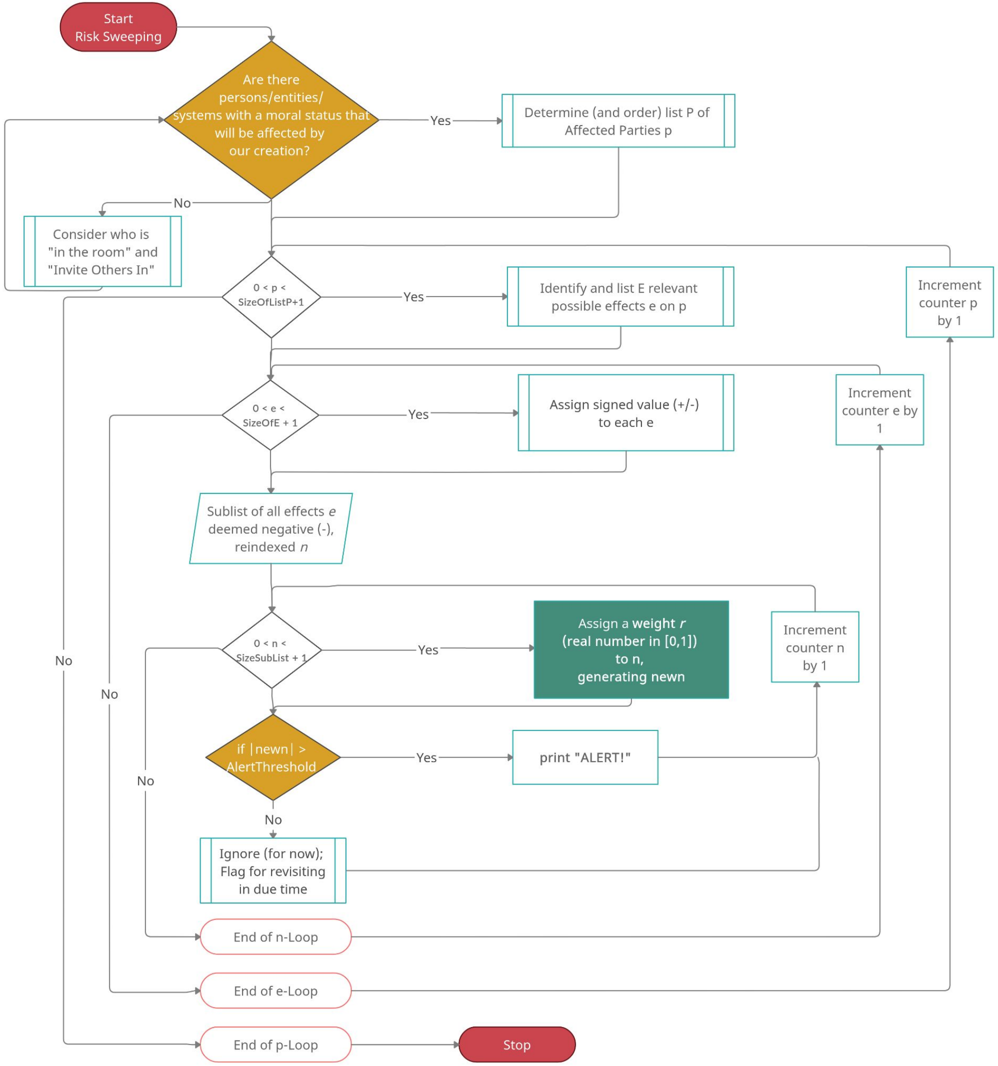
- Just as "risk sweeping" is a standard tool of good cybersecurity practice, **ethical risk sweeping** – where we inspect a system or creation for ethical risks and are "on alert" for particularly acute risks – is an essential tool for ethically sound design and engineering practice.
- Again, there exist a variety of systems or theories for assessing the rightness/wrongness of something, what counts as a "harm," and what other ethical risks may arise in any endeavor. We need not get into the weeds here about any of these approaches, or argue over the "best" one to use. Happily, there is enough overlap to guide us.
- Let's start with a particularly simple (and somewhat simplistic) walk-through of how one might evaluate rightness/wrongness (specifically, isolating harmful effects, and alerting one to ethical risks) – that is loosely Utilitarian in spirit. This is sketched in the following flow chart, which gives some guidance on how one might begin to approach ethical risk sweeping in a more regimented manner.

Routines for Addressing Ethical Considerations

1. Flow Chart for **Ethical Risk Sweeping**
   a. **Assign a Weight** Subprocess

2. **Ethics Checks:** Routines and Considerations
   a. Flow Chart for **Red Teams** (*Assume You Missed Risks)* Process
      i. **Determine Most Likely Causes** Subprocess
   b. Flow Chart for **Study Relevant Cases** Process
      i. - iv. Unpacking **Identify Parallels/Evaluate Choices Made**
   c. Flow Chart for **Thinking about Bad Actors** Process

1. Flow Chart for **Ethical Risk Sweeping**



Start Risk Sweeping

Are there persons/entities/ systems with a moral status that will be affected by our creation?

Yes

Determine (and order) list P of Affected Parties p

No

Consider who is "in the room" and "Invite Others In"

0 < p < SizeOfListP+1

Yes

Identify and list E relevant possible effects e on p

Increment counter p by 1

0 < e < SizeOfE + 1

Yes

Assign signed value (+/-) to each e

Increment counter e by 1

Sublist of all effects *e* deemed negative (-), reindexed *n*

0 < n < SizeSubList + 1

Yes

Assign a weight *r* (real number in [0,1]) to n, generating newn

Increment counter n by 1

if |newn| > AlertThreshold

Yes

print "ALERT!"

No

Ignore (for now); Flag for revisiting in due time

No

No

No

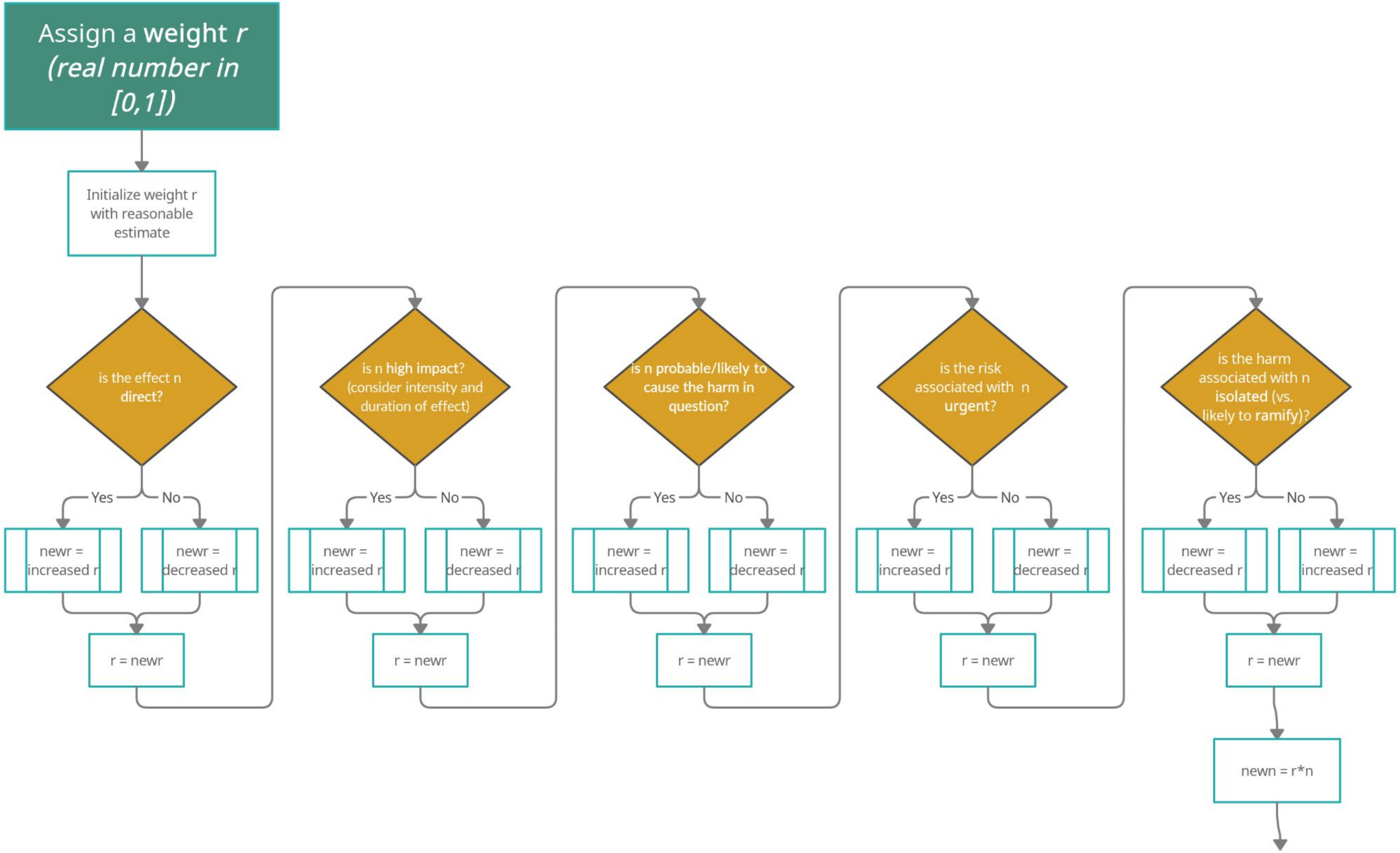
End of n-Loop

End of e-Loop

End of p-Loop

Stop

1a. Assign a Weight Subprocess



Assign a **weight** ***r*** *(real number in [0,1])*

Initialize weight r with reasonable estimate

is the effect n **direct?**

Yes

No

newr = increased r

newr = decreased r

r = newr

is n **high impact?** (consider intensity and duration of effect)

Yes

No

newr = increased r

newr = decreased r

r = newr

is n **probable/likely to cause the harm in question?**

Yes

No

newr = increased r

newr = decreased r

r = newr

is the risk associated with n **urgent?**

Yes

No

newr = increased r

newr = decreased r

r = newr

is the harm associated with n **isolated** (vs. likely to ramify)?

Yes

No

newr = decreased r

newr = increased r

r = newr

newn = r*n

Note on a Subtlety in Combining Weights

In the assessment of ethical risks, the *weighing* of a potential risk can be a rather subtle matter, especially if we were to fully respect the more nuanced interactions the weights themselves can have. For instance, consider that:

- Some risks may be low impact, but high urgency
- Some risks may be low urgency, but are simply too high impact to ignore
- Some risks may be high probability but low impact and isolated in their effects
- Some risks may be high impact, the result likely to ramify, while being low probability
- And so on...

Note on a Subtlety: Not all weights are created equal!

It's hardly ever as simple as saying that, for instance, 3 "increased weights" will amount to a greater total weight assigned than would an event that had 2 "increased weights."

For instance, consider certain risks associated with

> ***Event A***: *the creation of a virus which, if it got loose, would make all humans infertile, eventually ending our species.*

Such an event is high impact and highly non-isolated (ramified), but indirect, low probability, and the risk non-urgent. (2 increased weights, 3 decreased.)

Yet Event A is clearly of greater import to us than the risk associated with

> ***Event B***: *The creation of an (otherwise harmless) dietary supplement, soon to be released on the market, that contained caffeine without advertising this -- which, with high probability, would negatively affect the sleep of some of those who take it.*

Event B is high probability, urgent, mostly direct, but low impact and mostly isolated. (3 increased weights, 2 decreased.)

Note on a Subtlety in Combining Weights

- While it may at first seem unduly complicated if you were forced to write out explicitly and formally a total weight, we frequently have no problem making (and agreeing upon) determinations that, for instance, certain risks associated with events like Event A are of greater import than events like Event B.
- For now, it is okay to use your best informal judgment to figure out how to balance such combined weights and end up with a reasonable overall measure to determine whether or not the harmful effects of some aspect(s) of our creation are worth considering as important ethical risks, as compared to other candidates for our concern.
- However, remember to care about risks to others as much as you would a risk to yourself! There’s a danger in defaulting to weighing concerns that affect you or your in-group as if they were more important than the concerns of others.

Where Ethical Risk-Sweeping Can Go Wrong

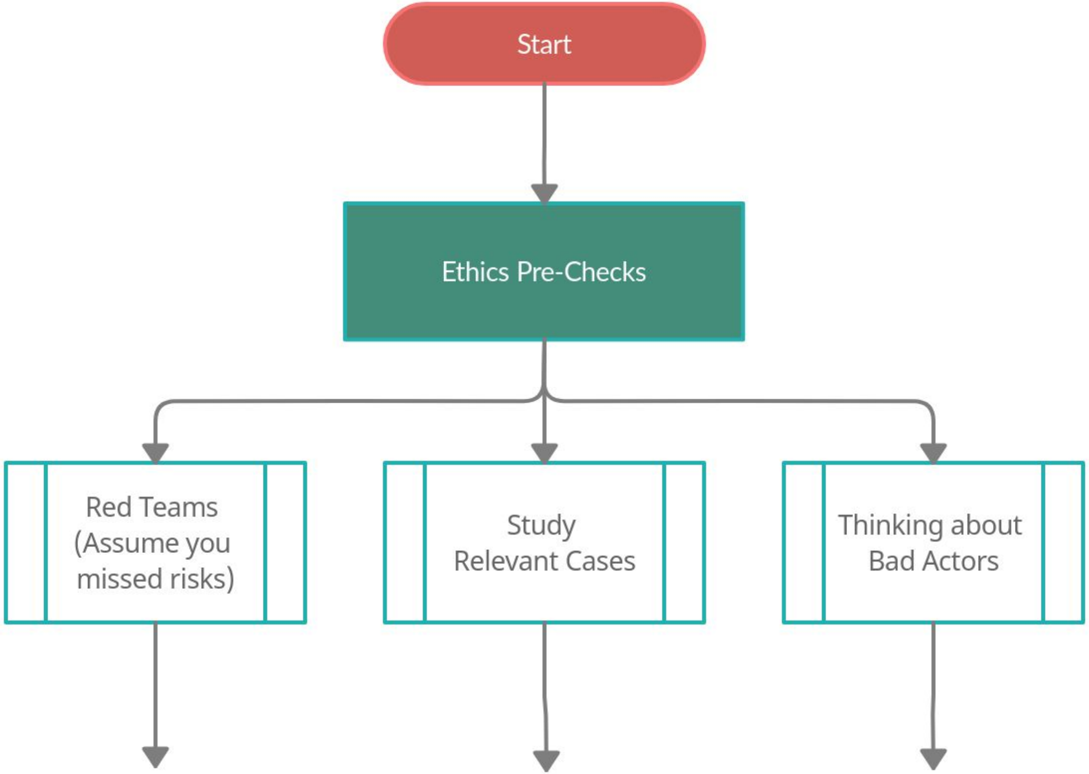
Ethical risk-sweeping can run into problems in a number of ways, and for a number of reasons. But some of the more conspicuous reasons for its breakdown may include:

- Our analysis of the causal interactions that might lead to harm is incomplete, inaccurate, or biased so that we are blind to the full scope of the harm
- We do not share key assumptions with certain of the people affected by our creation
- We do not think enough about the various ways our creation can be misused and abused, even if we ourselves have had good intentions in creating it
- The ethical risks are subtle and we oversimplify them in a way that is not productive or makes us blind to certain effects

Going Further: Preparing for Ethical Risks with Ethics Checks

- While the risk-sweeping protocol focuses on isolating individual risks, we can also devote attention to avoiding *systemic* ethical failures of a project. In order to deal with the various ways ethical risk-sweeping can go wrong, we should pair basic risk-sweeping with ways of inspecting the dynamics of systemic ethical failures.
- In technical design and engineering settings, we can use a variety of ***pre-check*** protocols or processes, to help address these concerns.
- The overall idea of this? Instead of *waiting* for ethical disasters to happen and then reacting to them, you should make use of tools for anticipating how larger-scale ethical failures of the project or creation *might* happen, and aim to understand, in advance, the preventable causes, and concrete alternatives, so that harms can be avoided or eliminated.

2. Ethics Pre-Checks: Routines and Considerations


Start
Ethics Pre-Checks
Red Teams (Assume you missed risks)
Study Relevant Cases
Thinking about Bad Actors

Ethics Pre-Checks: Important Routines

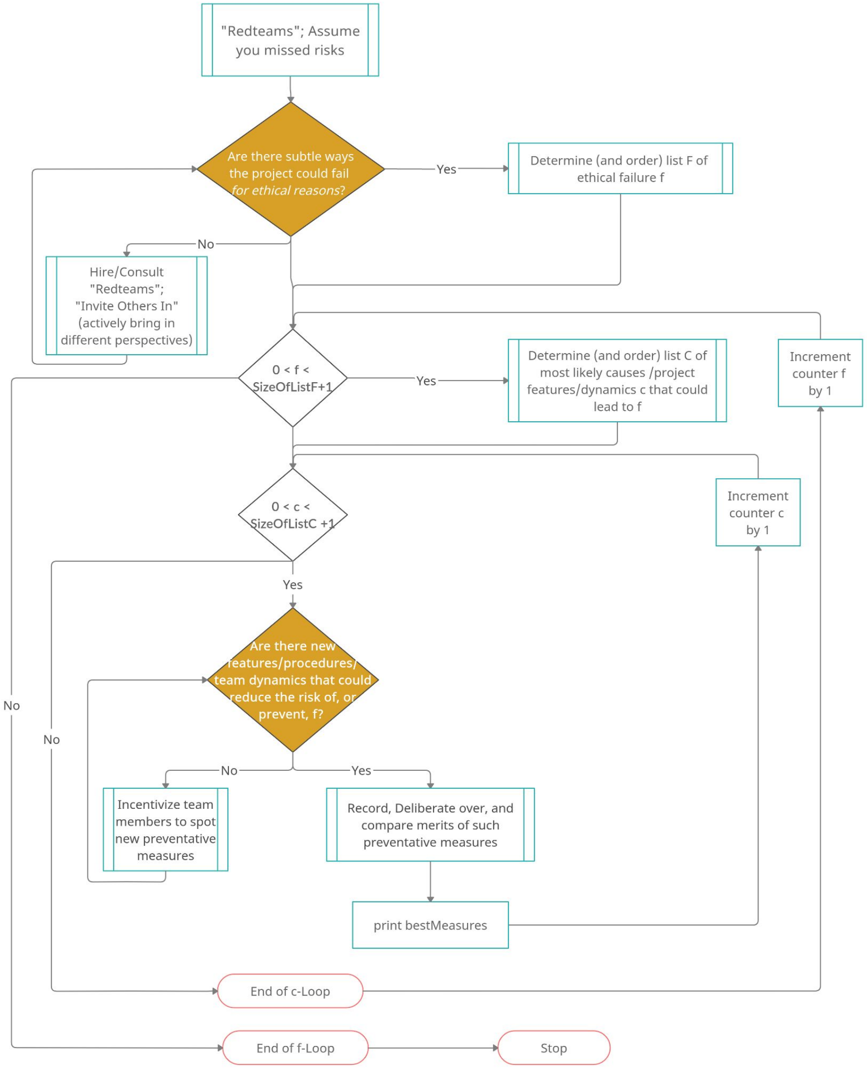
The three processes can help supplement risk-sweeping and address the ways it can go wrong. Specifically,

- Where our analysis of the causal interactions that might lead to harm is incomplete, inaccurate, or biased so that we are blind to the full scope of the harm – **Red Teams** (and other related tasks like **Inviting Others In**) can help us fill out the picture
- Where we do not share key assumptions with certain of the people affected by our creation – tools like **Inviting Others In**, and **Studying Relevant Cases**, can help us
- Where we do not think enough about the various ways our creation can be misused and abused, even if we ourselves have good intentions – tools like **Thinking About Bad Actors** will help us stay honest about these realities
- Where the ethical risks are subtle and we oversimplify them in a way that is not productive – using **Red Teams** (and assuming you missed risks) and **Studying Relevant Cases** will help identify some of these subtler risks.

Ethics Pre-Checks: Red Teams

- A **Red Team** is a group – typically used in cybersecurity and military/intelligence contexts – that adopts the role of an enemy or competitor, in order to provide security feedback.
- A Red Team assessment is similar to a penetration test in cybersecurity contexts, but it is *more targeted* – the aim is to test the detection and response capabilities of the original team/organization to a surprise attack or unexpected exploit. "Red Teaming," more generally, is any structured adversarial activity designed to uncover risks and vulnerabilities that may have been missed by conventional tests for vulnerabilities.
- Red Teams can be used to help uncover unexpected *ethical risks* as well, and one should seek out the feedback of those capable of providing this sort of ethically-oriented "red teaming" activity.

2a. Flow Chart for Red Teams (*Assume You Missed Risks*) Process



"Redteams"; Assume you missed risks

Are there subtle ways the project could fail *for ethical reasons?*

Yes

Determine (and order) list F of ethical failure f

No

Hire/Consult "Redteams"; "Invite Others In" (actively bring in different perspectives)

0 < f < SizeOfListF+1

Yes

Determine (and order) list C of most likely causes /project features/dynamics c that could lead to f

Increment counter f by 1

0 < c < SizeOfListC +1

Increment counter c by 1

Yes

Are there new features/procedures/ team dynamics that could reduce the risk of, or prevent, f?

No

No

No

Incentivize team members to spot new preventative measures

Yes

Record, Deliberate over, and compare merits of such preventative measures

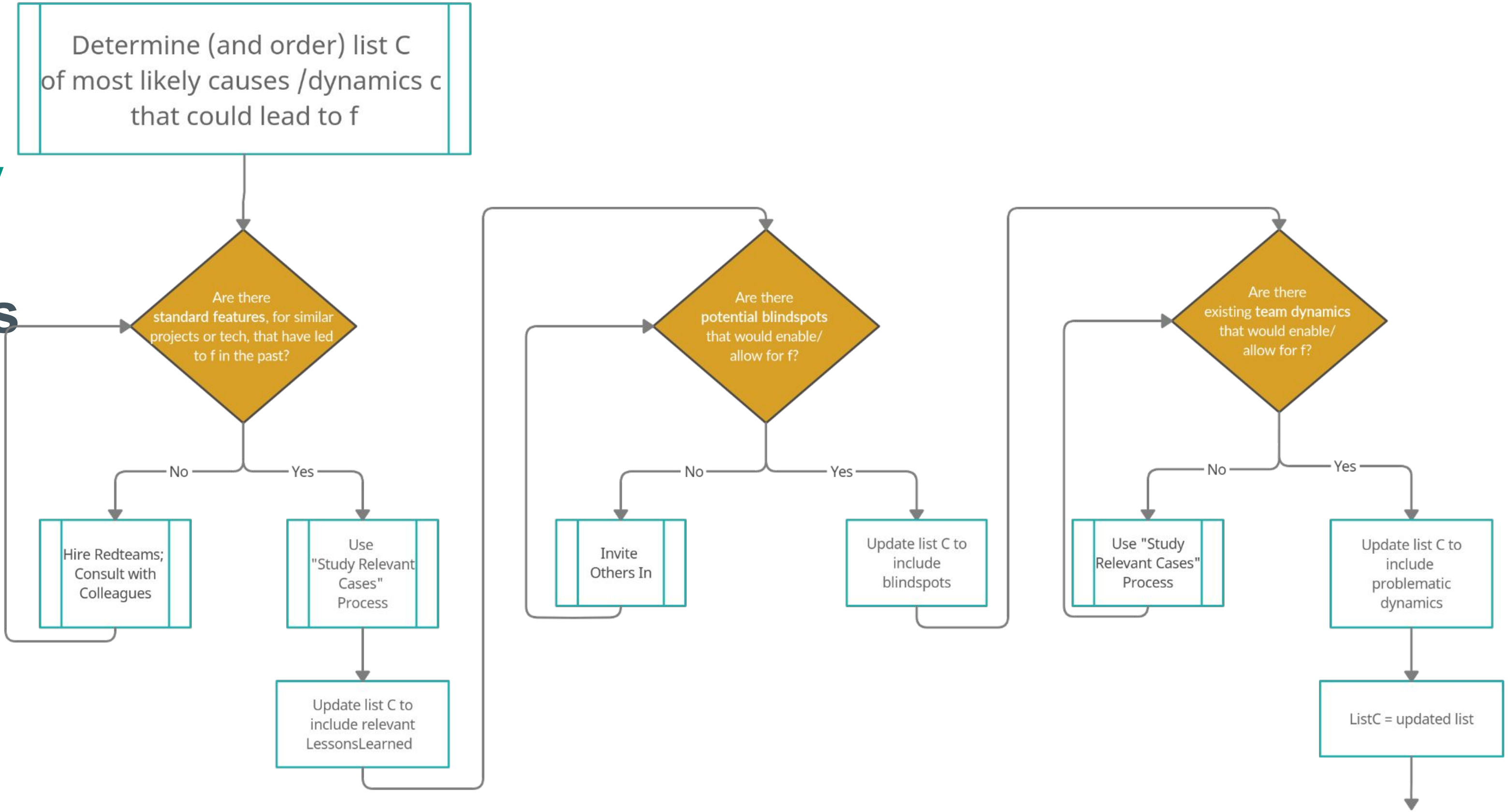
print bestMeasures

End of c-Loop

End of f-Loop

Stop

2a(i). Determine Most Likely Causes Subprocess



Determine (and order) list C of most likely causes /dynamics c that could lead to f

Are there **standard features**, for similar projects or tech, that have led to f in the past?

Hire Redteams; Consult with Colleagues

Use "Study Relevant Cases" Process

Update list C to include relevant LessonsLearned

Are there **potential blindspots** that would enable/ allow for f?

Invite Others In

Update list C to include blindspots

Are there existing **team dynamics** that would enable/ allow for f?

Use "Study Relevant Cases" Process

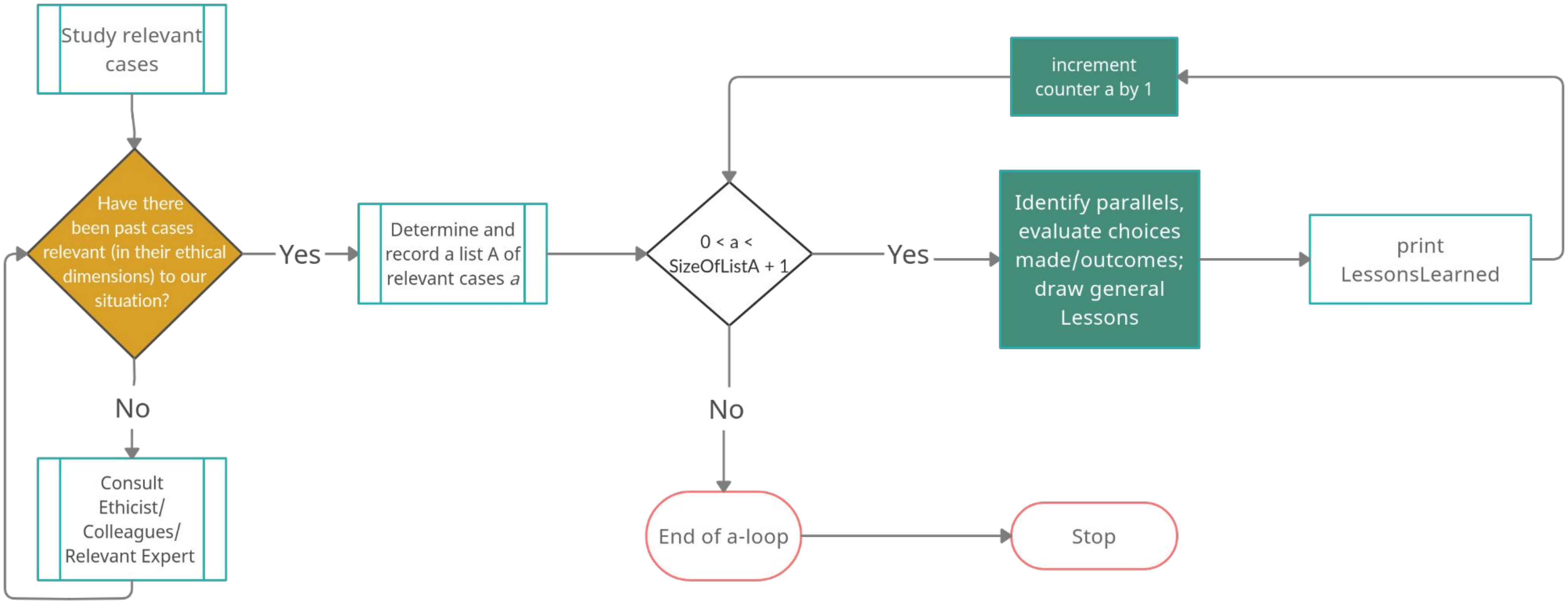
Update list C to include problematic dynamics

ListC = updated list

Ethics Pre-Checks: Studying Relevant Cases

- Another tool for helping us isolate and minimize ethical risks is **Studying Relevant Cases**.
- The procedure of **case-based analysis** is a more “bottom-up” approach, as opposed to a more “top-down” approach of applying a universal principle or ethical framework to a particular problem.
- Case-based analysis effectively is a procedure for helping to determine whether an action/decision is right or wrong by comparing the action/decision with unambiguous paradigm cases or closely related cases that have already been unpacked and analyzed.

2b. Flow Chart for Study Relevant Cases Process


Study relevant cases
Have there been past cases relevant (in their ethical dimensions) to our situation?
Yes
Determine and record a list A of relevant cases *a*
0 < a < SizeOfListA + 1
Yes
Identify parallels, evaluate choices made/outcomes; draw general Lessons
print LessonsLearned
increment counter a by 1
No
Consult Ethicist/ Colleagues/ Relevant Expert
No
End of a-loop
Stop

2b(i-iv). Unpacking Identify Parallels, Evaluate Choices Made...

In more detail, the procedure of case-based analysis involves:

i. **Identifying Paradigm Cases**

ii. **Identifying Relevant Parallels/Differences**

iii. **Evaluate Choices Made/Outcomes of the Paradigm Cases**

iv. **Using Analysis as a Guide**

2b(i). Identifying Paradigm Cases

- **General task:** Identify similar or paradigm cases that include clearly right/wrong decisions and mirror the present case in certain respects
- **Questions to ask:**
  - ❑ Where/when has a case similar (in its ethical dimensions) to our case occurred before?
  - ❑ Which of these are ‘paradigm’ cases of the kind of ethical situation facing us? For instance, when concerned about a particular ethical risk, identify an appropriate negative paradigm case, where a clearly wrong action or negative outcome occurred.
- Things to do:
  - ❑ Extract the relevant features or aspects of the situation that contributed to making the activity/decision right or wrong.

2b(ii). Identifying Relevant Parallels/Differences

- **General task:** Identify relevant parallels between, and/or differences among all the cases, including the present case
- **Questions to ask:**
  - ❑ What are some notable similarities/differences across the cases?
  - ❑ In what ethically relevant respects is the present case similar to these paradigm cases? For example, the test case may affect the same group of stakeholders, or introduce the same risks, or present fundamentally the same moral dilemma.
  - ❑ In what ethically relevant respects is the present case different from the paradigm cases? For example, this time the affected parties may be very different or harder to pin down, or the stakes may be lower.

2b(iii). Evaluate Choices Made/Outcomes of the Paradigm Cases

- **General task**: Carry out an analysis – for instance, following Quinn* – and evaluate choices made/outcomes; and/or make use of prior similar analyses, similar to the way a judge might base a decision on a prior court-ruling

- **Questions to ask:**

  ❑ What was done in the paradigm cases? What choice was made, how was the dilemma resolved, and how was the decision justified?

  ❑ What was the outcome? Who benefited, and how? Who got hurt, and how? How did the public/media/regulators/other relevant agents react to the choices made? How did they respond to the subsequent justifications given?

  ❑ Did those who made those choices come to regret them, or renounce them, or did they continue to defend them?

  ❑ Did they institute any safeguards or changes, in the aftermath of the decision?

  ❑ In what ways does this provide a template or model of ethical success; and in what ways does it function as a warning?

*See Quinn: https://dl.acm.org/doi/pdf/10.1145/1124706.1121503

2b(iv). Using Analysis as a Guide

- **General task**: Use the analysis and analogical reasoning to isolate lessons, opportunities, solutions/decisions, and preventative protocols for the test case.

  This effectively concerns determining how our ethical knowledge of the paradigm cases ought to influence our ethical reasoning and judgment in *this* case.

- **Questions to ask**:
  - ❑ What lessons should transfer over?
  - ❑ What solutions that worked well before are likely to work well again?
  - ❑ What mistakes that they made in the paradigm case are we in danger of making *now*?
  - ❑ What risks that were successfully mitigated *that* time can be mitigated with similar strategies now? And are there any mitigation strategies that are no longer appropriate?
  - ❑ How else do the relevant differences between the cases limit or alter the transferability of the paradigm lessons to this present case?

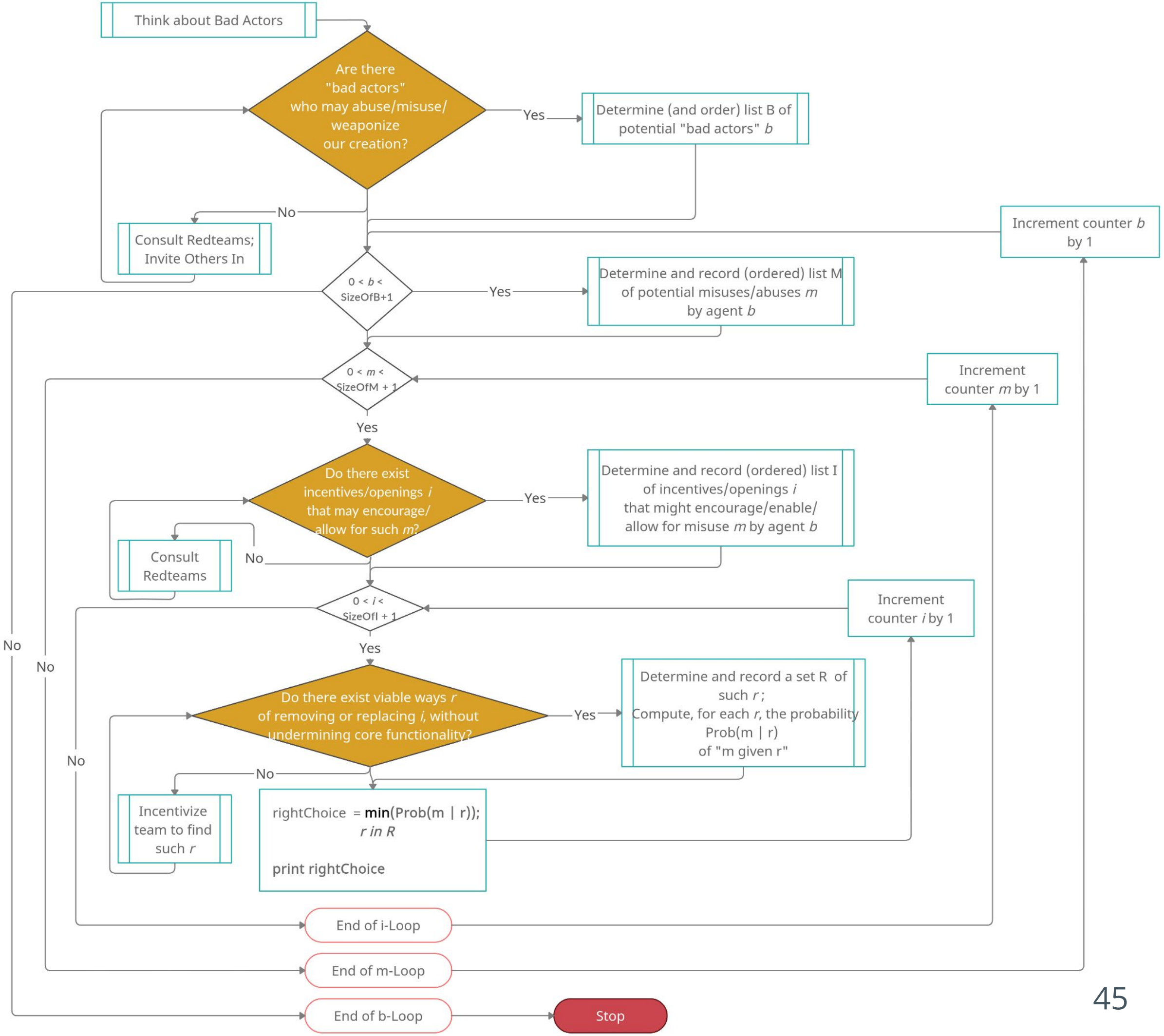
Ethics Pre-Checks: Thinking about Bad Actors

- Any technology is a form of power – and there will be those who may use such power to their own advantage, in ways that harm others. There will be others who may use such power in remarkably thoughtless ways.

- We can use the term **Bad Actors** to include:
  - ❑ anyone who may want to abuse or weaponize your creation
  - ❑ anyone who will misuse or misinterpret it with extreme irrationality
  - ❑ anyone who otherwise has some bad intentions with respect to your creation.

- In building powerful tools, it is part of your responsibility that you anticipate and mitigate the abuses/misuses of such bad actors.

2c. Flow Chart for Thinking about Bad Actors Process



Think about Bad Actors

Are there "bad actors" who may abuse/misuse/ weaponize our creation?

Determine (and order) list B of potential "bad actors" *b*

Consult Redteams; Invite Others In

Increment counter *b* by 1

0 < *b* < SizeOfB+1

Determine and record (ordered) list M of potential misuses/abuses *m* by agent *b*

0 < *m* < SizeOfM + 1

Increment counter *m* by 1

Do there exist incentives/openings *i* that may encourage/ allow for such *m*?

Determine and record (ordered) list I of incentives/openings *i* that might encourage/enable/ allow for misuse *m* by agent *b*

Consult Redteams

0 < *i* < SizeOfI + 1

Increment counter *i* by 1

Do there exist viable ways *r* of removing or replacing *i*, without undermining core functionality?

Determine and record a set R of such *r*; Compute, for each *r*, the probability Prob(m | r) of "m given r"

Incentivize team to find such *r*

rightChoice = **min**(Prob(m | r)); *r in R*

print rightChoice

End of i-Loop

End of m-Loop

End of b-Loop

Stop

CASE WESTERN RESERVE UNIVERSITY

INAMORI INTERNATIONAL CENTER FOR ETHICS AND EXCELLENCE

3. In Review:
Looking Back to Look Forward

Consider this: Feedback and Iterating for Improvement

- **Remember That Ethical Design/Engineering Is Never a Finished Task**

Don’t just “check off” the ethical issues and associated demands, expecting only to return to them in the event of a disaster. Make reflection on these matters an ongoing practice – one that has a clear notion of what “progress” (and what “failure”) looks like.

Consider this: Feedback and Iterating for Improvement

- **Identify Feedback Channels that Will Deliver Reliable Data on Ethical Impact**

There’s no way to know whether, or to what extent, efforts to mitigate or address ethical risks are actually succeeding in that fullest sense unless we regularly gather reliable data (and conduct “post-checks”) on the ethical impact of our designs on society and specific stakeholders.

Where does that data come from? Well, it won’t come at all unless instruments are specifically designed to elicit and transmit feedback of that kind. Thus, any design plan should identify specific instruments/input channels by which ethical impact data will be collected and audited – from users, but also from other affected stakeholders (including systems like the environment).

Consider this: Feedback and Iterating for Improvement

- ❑ **Develop Formal Procedures and Chains of Responsibility for Ethical Iteration**

The auditing of ethical impact cannot be an ad hoc, occasional event; it must become a standard feature, one that is checked and updated regularly.

- Make sure to consider how feedback from such “ethical audits” and “post-checks” will get analyzed, communicated, and used in the next iteration of the creation or design cycle.

- Who will be accountable for closing the loop? Who is accountable for ethical design overall?

Consider this: Feedback and Iterating for Improvement

- ❑ **Develop Formal Procedures and Chains of Responsibility for Ethical Iteration**

Remember that these things won’t happen on their own. A system often gets whatever its incentive structures and chains of responsibility reward.

- If done right, formal procedures and chains of responsibility for ethical design and engineering help create and support the incentives that drive the right sort of behavior;
- if done wrongly, no amount of “good intentions” can save a team from ethical disaster.

Consider this: No One is Perfect!

We all make mistakes. We all make inadvertent errors. We all transgress rules or fail to uphold values. What to do when you something wrong?

- “Don’t Panic!” – Douglas Adams. The cover-up is nearly always worse than the crime. The first, best thing you can do is own up to what you did and don’t try to hide it or blame someone else.
- Focus on determining exactly where and why things went wrong. Try not to rationalize. Describe the situation honestly to yourself first. (“I did X, because Y.”)
- Offer to share your story, and your analysis of the situation, so that your mistake can give everyone in the engineering community an opportunity to recommit themselves to core values and ideals. Even if you can no longer undo the wrong, the rest of your community can learn from your mistakes and draw important lessons. Help them learn the right lessons by being open about what went wrong.

Consider this: The Power of Shared Values

While it is an important task to scan for and address new ethical risks and tp be vigilant to any potential oversights/blindspots/failures that may lead to harms, it also pays to institute procedures for reflecting on and reinforcing *positive goods* sought and the *shared values* driving you and those you work with.

- Oftentimes, short-term benefits or narrow goals can lead us to lose sight of what motivated us to do the work in the first place. In the extreme, such over-focus on short-term incentives/goals can cause us to undermine, pervert, or sacrifice the bigger goals (and the values these are motivated by).

- It is important to frequently find time to "check in" with oneself and co-workers about these bigger-picture values – and honestly assess whether the work you are doing is advancing those aims, or whether you have started to sacrifice those aims.

Exercise

See Attached Sheet

Further Readings and Resources

- Quinn, Michael J. *Ethics for the Information Age*. Boston: Pearson/Addison-Wesley, 2020. Print.This book is a nice resource for the student looking to delve deeper into some of the issues covered in these modules, and related topics facing ethics of technology. A number of the examples reviewed in the first module were adapted from this text, which includes more extensive discussion of some of these.

- Markkula Center for Applied Ethics: Ethics in Technology Practice
  The Markkula Center is an excellent resource for materials on ethics of technology and other applied ethics areas, offering thousands of pages of resources on applied ethics (including case studies, links, podcasts, videos, and more). Elements of this module build on these resources.

Further Readings and Resources

- Shannon Vallor's book *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting* provides an excellent overview of what "we need to live wisely and well with emerging technologies."
- Safiya Umoja Noble's book *Algorithms of Oppression* and Ruha Benjamin's book *Race After Technology* are good places to start to learn more about the cause and effects of bias in algorithms.
- A curated list of some good resources on Ethical Issues in AI: https://www.aiethicist.org/ethics-cases
- *Awful AI* is a curated list to track current harmful usages of AI: https://github.com/daviddao/awful-ai