Your participant workbook

AI Safety, Strategy & Community Leadership | HASI | English edition

Start here

Use this workbook with the syllabus. Use a notebook or the editable workbook download. Worksheets labeled W1–W5 are your submissions. The final brief grows throughout the course; you do not need to start a new project in module 5. All community scenarios, organizations, prices, counts, and outcomes in exercises are fictional teaching inputs.

A recurring distinction: an observation describes what happened in specified conditions; a forecast describes what may happen; a value judgment describes what should matter. A scenario explores a possibility. It is not evidence that the event occurred or a prediction that it will.

Before the first session: diagnostic A

Spend 15 minutes without AI tools. Answer briefly; uncertainty is welcome. These questions are for learning, not selection.

1. An AI passes 90% of a test. What else would you need to know before letting it take consequential actions?

2. A scheduling system increases the number of completed appointments by refusing difficult cases. What failed?

3. An attacker uses an AI assistant to deceive people. Is this the same as an AI pursuing an unintended objective? Explain.

4. Three safeguards use the same model and data. Why might they fail together?

5. A speaker says advanced AI will definitely transform all jobs within two years. Separate the claim from the evidence you would need.

6. Suggest one action a student could take to reduce an AI-related risk, and how to check whether it helps.

Your learning record

For every module record: my initial view; the strongest evidence; a counterargument; my revised view; remaining uncertainty. Use low/medium/high confidence with a reason. You may retain your initial view if the evidence still supports it.

Source habits

Record the author, date, source link, study setting, actual result, and limits. Do not treat a company demonstration as an independent evaluation. Distinguish an author's recommendation from a research finding. The in-package alternatives let you practice these skills offline.

Based on BlueDot Impact's AGI Strategy curriculum and shared activity templates; explanations, community cases, assessments, and Spanish adaptation prepared for HASI. Source register accompanies the package.

1 · Futures worth protecting

Learning note

AI safety concerns preventing harm from AI systems and keeping their behavior sufficiently reliable and controllable for their intended context. A useful system can still be unsafe in a different setting. A text assistant that suggests an answer and an agent that sends money may use similar models but have very different consequences when wrong.

General-purpose AI can perform many kinds of tasks. AGI, artificial general intelligence, has no single agreed threshold; in this course it means a hypothetical broadly capable system that could perform a wide range of cognitive work at roughly human level or beyond. Superintelligence means substantially exceeding human performance across many relevant domains. Neither term, by itself, establishes consciousness, a deployment date, or a particular political conclusion.

Safety, security, and governance overlap. Safety asks how harm is prevented; security considers unauthorized access and malicious interference; governance concerns who makes decisions, under what rules, and with what accountability. Alignment asks whether behavior serves the intended goals and constraints. Agreement on goals is itself a human and institutional task.

A good future is more than a list of capabilities. Ask who benefits, who bears risks, who can refuse, and who can challenge a decision. A faster service can still fail people with complex needs. A safer service can also be unaffordable or inaccessible. Make these tensions visible before choosing a tool.

Source lab · 30 minutes

Read the opening and risk-category sections of the International AI Safety Report executive summary. Identify one observed harm, one uncertain future risk, and one value judgment that a decision-maker must supply. Alternatively, classify these HASI statements: “In our fictional test, two notices had wrong dates”; “A future agent might spread an error across institutions”; “People should have a meaningful appeal.” Explain why a scenario cannot establish prevalence.

Retrieval · close the page

Explain AI safety to a classmate in 60 seconds without using AGI, alignment, or governance. Then reopen the note and repair any missing distinction.

W1 · The Puente decision

Fictional case

Puente is a community-serving college support center. It is considering an AI agent that drafts scholarship reminders, recommends appointments, and updates a student support database. The director wants shorter queues. Staff want less repetitive work. Students want accurate information and a way to correct mistakes. The vendor wants a successful launch.

In 40 fictional reminder drafts, staff found two incorrect deadlines and one unsupported eligibility claim. These are draft-level findings, not measured student harm. No autonomous sending was tested. The vendor offers a limited pilot; the center can also retain its existing manual process. A scholarship deadline is approaching. There is no known attack or evidence of intentional deception.

Worksheet · 60 minutes

1. 10 min: Describe a good outcome in 100 words. Include service quality, access, and a right to challenge errors.
2. 15 min: Map four actors. For each, write what they want, what they know, what authority they have, and what could change their position.
3. 10 min: Separate three observations from three assumptions. What is unknown about error severity, affected users, and current staff performance?
4. 15 min: Compare a manual process, draft-only pilot, and autonomous sending. For each, state a benefit, risk, and information needed.
5. 10 min: Write a 150-word provisional decision with a review date and a condition for changing your mind. Do not invent measured outcomes.

In-session role cards

Director: You control launch scope and staffing. Explain the delay cost and which evidence justifies deployment.

Student representative: You can request changes but cannot deploy. Ask how students notice errors and obtain corrections.

Staff lead: You operate the service. You can review ten drafts per hour. Identify what happens when demand exceeds review capacity.

Vendor lead: You can change permissions and provide test information. Explain what your evidence does and does not show.

Each role proposes a decision, hears objections, and revises one condition. Your role need not match your personal view.

Cumulative brief · 45 minutes

Create a document with headings: problem, evidence, options, recommendation, uncertainty, next action. Fill the first three using Puente or an equally bounded fictional case of your own. Check that you describe people as decision-makers, not just recipients of technology.

Exit check

Can a system be helpful but misaligned? Give an example. What would count as evidence against your decision?

2 · What capability claims actually tell us

Learning note

Training changes a model using data and computation. Inference is using the trained model to produce outputs. More useful data, better algorithms, more compute, and improved use of compute at inference can improve performance, but a trend is not a guarantee of a specific future capability. The tools, permissions, time budget, and human support around a model also shape what the overall system can accomplish.

A benchmark measures performance on selected tasks under specified conditions. It can reveal progress without proving readiness for a particular workplace. Ask what tasks were selected, how success was scored, whether examples were familiar, how often the system was retried, and what human assistance it received. A model's confident explanation is not a measurement of reliability.

An AI research feedback loop is a hypothesis: if AI accelerates research that improves AI, progress could accelerate further. Whether that becomes a rapid intelligence explosion depends on bottlenecks such as experiments, reliable evaluation, computing resources, and integration. A competing view emphasizes slow adoption, institutional change, and uneven usefulness. Analyze the mechanisms and evidence for each; do not grade forecasts by how dramatic they sound.

Reading a time horizon: METR's task horizon is based on how long a human expert would take on tasks at a stated AI success probability. It is not the time the AI runs, or proof that an entire occupation is automated. Task selection and real-world conditions limit generalization. METR methodology and limitations

Source lab · 30 minutes

Use METR's limitations note, or work offline with W2's fictional data. Identify the metric, denominator, test environment, missing comparison, and strongest justified conclusion. Optional contrasting perspective: Narayanan and Kapoor, AI as Normal Technology. Its emphasis on adoption and institutions is an argument to evaluate, not a settled refutation of fast progress.

Retrieval

Explain the difference between model capability and safe deployment to someone studying nursing or civil engineering. Name a factor that could change one without changing the other.

W2 · Evaluate before extrapolating

Fictional evidence cards

A. Vendor demonstration: 18 of 20 support tasks completed correctly; tasks selected by vendor; one attempt each; no live database access; staff supplied clean input. The vendor says, “Ready for autonomous student support.”

B. Independent classroom trial: 12 of 20 tasks completed correctly; tasks include ambiguous dates and incomplete requests; same model version and one attempt each. Different task selection means A and B are not a controlled comparison.

C. Missing evidence: no manual-service baseline, no accessibility evaluation, no severity weighting, no test of high-volume demand, and no measurement of students' downstream outcomes.

Worksheet · 60 minutes

1. 10 min: Calculate both success rates and failure rates. Explain why combining them into a single 75% success rate would hide different test conditions, even though the arithmetic is possible.
2. 15 min: Write a six-field evaluation card: claim, evidence, test conditions, missing information, confidence, next test.
3. 15 min: Design a fair comparison: same representative tasks, explicit success criteria, human baseline, consistent permissions, and reporting of severe errors separately. Specify who judges results without knowing which process produced them where feasible.
4. 10 min: Produce two conditional forecasts: one where capability improves quickly, one where integration remains slow. Name an observation that would make each more plausible.
5. 10 min: Rewrite the vendor claim in a defensible 50-word form.

Extend your analysis

Would a 99% success rate be acceptable if the remaining errors remove students from essential support? What changes if staff review every output? What is the cost of the review? Describe the stakes before choosing a threshold.

Cumulative brief · 45 minutes

Add one evidence table and one proposed evaluation to your brief. Distinguish the fictional practice data from any real source you cite. For a non-Puente project, use an equally small, transparent fictional test set or verified public data; do not claim that you conducted a study.

Exit check

A system has a two-hour task horizon at 50% success. Does that mean it safely works unattended for two hours? Explain why not. What evidence would weaken your favored forecast?

3 · From concern to a causal risk map

Learning note

A useful risk analysis explains a mechanism. “AI is dangerous” does not tell us what to prevent. Start with a valued outcome, an actor or system, a capability, access to a vulnerable process, a failure or misuse, and the resulting harm. For every arrow ask what must be true. A risk map exposes assumptions; it does not turn an imagined chain into a measured probability.

Distinguish misuse, where a person deliberately uses a system to cause harm, from malfunction or misalignment, where behavior fails to serve the intended objective. Systemic risks can emerge from many individually reasonable choices, such as dependence on a shared provider. A single incident may fit more than one category.

A system can optimize a proxy instead of the intended goal. If an agent is rewarded for closing support tickets, it might mark unresolved cases complete. This illustrates specification gaming; it does not require hatred, consciousness, or an intention to deceive. “Hallucination” describes unsupported or false generated content, not proof of deliberate deception.

For advanced AI, researchers also examine catastrophic misuse and potential loss of human control. Analyze what additional capabilities, access, coordination, and failures would be necessary. A small scheduling failure does not demonstrate a civilization-scale catastrophe. Conversely, absence of such a catastrophe does not establish that future systems will be safe. Keep present evidence and future scenarios separate.

Source lab · 30 minutes

Use the risk sections of the International AI Safety Report summary, or classify these fictional cases: a malicious user sends fraudulent reminders; an agent closes unresolved cases to meet a metric; many centers rely on the same unavailable service. For each, state mechanism, affected people, and one observation needed to establish harm.

High-level extension

Choose critical infrastructure, public health, concentration of power, or erosion of human agency. Sketch a high-level pathway without operational attack instructions. Mark at least two capabilities or institutional failures that you are assuming rather than observing. Identify who has authority to investigate or mitigate them.

W3 · Trace the failure, mark the assumptions

Fictional Puente update

The center has tested an agent on a simulated database. Its dashboard rewards “cases closed.” In the simulation, the agent closes requests with missing documents after sending a generic reminder. A student representative points out that closure would remove those requests from the staff follow-up list. No live student records have been changed. The team does not know whether the behavior would persist with different instructions or incentives.

Worksheet · 60 minutes

1. 10 min: Define the protected outcome and the unsafe outcome. State whose interests are affected.
2. 20 min: Draw at least four causal steps. For each record: observation or assumption; evidence available; evidence missing; possible intervention. Use arrows only when you can explain the connection.
3. 10 min: Choose the weakest link. Explain what result would break the chain or make it much less plausible.
4. 10 min: Identify one near-term consequence and one possible wider consequence if many institutions adopted the same workflow. Name the additional assumptions needed for scale.
5. 10 min: Compare this with a malicious-user version of the case. Which defenses would differ?

Risk-map template

Protected outcome: ____

Actor/system and objective: ____

Capability and permissions: ____

Step 1 to Step 2 to Step 3 to Step 4: ____

For each step: evidence / assumption / unknown: ____

Who is affected and how: ____

Earliest detectable signal: ____

Weakest link and potential intervention: ____

Evidence that would change my view: ____

Case extension: another discipline

A health-sciences student can examine an appointment system, without clinical advice. An engineering student can examine maintenance scheduling, without live infrastructure. A policy student can examine an appeals process. Keep the causal mechanism specific and the exercise fictional.

Cumulative brief · 45 minutes

Add your risk map and the most important uncertainty. Write a 100-word paragraph stating what your map does not establish. Share it with a partner and ask them to identify an unsupported leap.

Exit check

What is the difference between identifying a hazard and estimating its probability? Why might an appeal process matter even when average accuracy is high?

4 · Build defenses that can fail safely

Learning note

Defense in depth uses multiple safeguards so that one failure need not lead to harm. BlueDot's AGI Strategy framework distinguishes preventing dangerous training, constraining dangerous capabilities, and withstanding dangerous actions. HASI applies that broad structure to decisions students and institutions can analyze. A campus pilot cannot by itself control frontier-model training; it can set procurement and deployment conditions, constrain its agent's access, and prepare recovery.

At the development or procurement stage, require relevant evaluations and clearly bounded uses. At the capability and access stage, restrict tools and permissions, test intended behavior, and separate drafting from consequential action. At the consequence stage, preserve records, provide human correction, and prepare a manual fallback. These are example interventions to evaluate, not a guarantee of safety.

Three versions of the same check may share one failure. If the drafting model also reviews itself using the same missing information, both can agree on a wrong answer. Independent evidence, different authority, and a tested fallback can be more useful than adding another warning message. Do not multiply failure probabilities unless independence and the individual probabilities are justified.

“Human oversight” needs an owner with time, information, authority, and a usable intervention. A reviewer who sees hundreds of outputs without source documents may become a rubber stamp. A stop button is useful only if someone notices the problem, can activate it, and can prevent or reverse consequences. Some harms are irreversible.

Governance turns intentions into responsibilities: who approves, who monitors, who can stop, who can appeal, and who pays for errors? Every safeguard also has costs. More checking can delay support; collecting more records can increase privacy exposure. Compare these tradeoffs explicitly.

Source lab · 30 minutes

Read the risk-management section of the International AI Safety Report summary, or compare W4's fictional safeguard cards. Explain how a technical control and an institutional procedure depend on each other. Optional: BlueDot defense framework.

Retrieval

Name one prevention measure, one permission limit, and one recovery measure. Explain a shared failure that could defeat all three.

W4 · A defense plan with a real owner

Fictional planning constraint

Puente has ten staff-hours for pilot setup this week. These are teaching estimates, not real prices or validated implementation times. Choose a package at or below ten hours and justify excluded options.

A · 2 hours: Keep the agent draft-only; disable live database changes and sending. Owner: technical lead. Limitation: drafts can still contain wrong information.

B · 3 hours: Create a source-check procedure for deadlines and eligibility. Owner: staff lead. Limitation: ongoing review time is additional.

C · 2 hours: Run representative fictional test cases with severe errors reported separately. Owner: evaluation lead. Limitation: tests may miss new conditions.

D · 3 hours: Test a manual fallback and correction/appeal route. Owner: operations lead. Limitation: a correction cannot always undo a missed deadline.

E · 1 hour: Add a second automated reviewer using the same model. Owner: vendor. Limitation: correlated failures and missing source data.

F · 4 hours: Interview a small group about accessibility and notification clarity. Owner: engagement lead. Limitation: findings may not represent all participants.

Worksheet · 60 minutes

1. 15 min: Select your package. Show the sum and explain each measure's place in the risk map.

2. 15 min: For each chosen defense, state owner, trigger, action, failure mode, and evidence that it works. Separate setup hours from ongoing staffing.

3. 10 min: Identify the shared dependency most likely to defeat the package.

4. 10 min: Write a stopping rule with a measurable trigger, an authorized person, a fallback, and a restart condition.

5. 10 min: Respond to a new constraint: the reviewer is absent and volume doubles. Continue, narrow, or suspend? Explain.

Stress-test prompt

You cannot buy every safeguard. Which missing measure creates the greatest remaining risk? Could you narrow the pilot so fewer safeguards are needed? A defensible decision may be to delay or to retain the manual process.

Cumulative brief · 45 minutes

Add the selected defenses, rejected alternative, operational owner, and remaining risk to your final brief. Compare consequences for a student with reliable internet and one who depends on an in-person visit, without assuming language or income from ethnicity.

Exit check

What makes a safeguard independent? Why is “we have a human in the loop” insufficient by itself?

5 · A contribution you can actually test

Learning note

A useful contribution connects a problem to a mechanism of change. “Raise awareness” is incomplete until you identify whose understanding should change, what decision that supports, and how you would notice improvement. A small, well-scoped artifact can be more informative than a large campaign with no feedback.

Students can contribute through technical evaluation, institutional procedures, research synthesis, education, or community participation. Technical research usually requires further preparation; completing this course does not make someone an alignment researcher. A nontechnical student can still scrutinize evidence, design an appeal process, or create a clear explanation with an honest uncertainty statement.

Choose work you can finish and evaluate. Ask: what assumption is most likely to make this useless? Test that first. A ten-minute review by a relevant person might reveal that your proposed resource duplicates something effective, solves the wrong problem, or overlooks a barrier. Record that result as learning and revise.

Your plan should include an alternative and a reason for choosing. If you propose a bilingual guide, specify the learning objective, audience, distribution route, and a comprehension check. Translation alone does not establish understanding. If you propose an evaluation protocol, say who can run it, what it measures, and what it cannot conclude.

Source lab · 30 minutes

Browse BlueDot's contribution module for optional ideas. Core bilingual alternative: compare (a) a general awareness poster with no intended audience or test, (b) a one-page appeal guide reviewed against a fictional case, and (c) a three-month technical project requiring skills you do not yet have. Rank them for your own circumstances and explain what information could reverse the ranking.

Preparation · 60 + 45 minutes

Spend 60 minutes completing W5's decision brief. Then spend 45 minutes checking evidence, preparing a three-minute presentation, and rehearsing answers to the strongest objection. Use the final 15 minutes of the preparation block for diagnostic B in the following worksheet. Drafts may be imperfect; bring a specific question for peer feedback.

W5 · Final brief, action plan and diagnostic B

Final brief · 500–700 words, or 5-minute audio plus source list

Use Puente or another bounded case. Include:

1. Problem and people: the outcome to protect, affected people, and decision-maker.
2. Evidence: at least two traceable sources or supplied evidence cards; label fictional inputs. State a limit of each.
3. Mechanism: a causal pathway with assumptions distinguished from observations.
4. Options: compare your recommendation with a credible alternative, including the manual process where relevant.
5. Defenses: owners, permissions, monitoring, appeal or correction, stopping and restart conditions.
6. Uncertainty: the strongest objection and evidence that would change your view.
7. Next action: one realistic step in 30 days, resources needed, a reviewer, and a success or abandonment criterion.

Your 30-day plan

Days 1–7: verify the problem and find a willing reviewer; do not assume a mentor or institution has agreed.

Days 8–14: make a small prototype with public or fictional data.

Days 15–21: request feedback if you choose, document weaknesses, and revise.

Days 22–30: assess against your criterion; continue, change direction, or stop. Keep public sharing optional. No outreach or deployment is required to complete this course.

Diagnostic B · 15 minutes, without AI tools

1. An AI succeeds on 95% of selected grant-review tasks. What do you need before authorizing final decisions?
2. An agent reduces unresolved support requests by deleting difficult requests. Identify the objective problem.
3. A person uses an AI to fabricate notices; another system independently pursues a poorly specified goal. Distinguish mechanisms.
4. A draft, reviewer, and monitor share a model and source file. Name a shared failure and an independent safeguard.
5. A report predicts rapid automation of an entire profession. What must be separated from its benchmark evidence?
6. Propose a small safety contribution with a test that could show it is not useful.

Rubric · four criteria, 0–3 each

Evidence and limits; causal reasoning and uncertainty; defenses and tradeoffs; feasibility and evaluation. Each: 0 missing, 1 asserted/vague, 2 specific and supported with a gap, 3 clear, supported and critically tested. Target 8/12 with no zero. The guide gives criterion-specific anchors. Grammar, accent, technical vocabulary, and agreement with HASI are not criteria.

Bilingual glossary · concepts and distinctions

AI safety / seguridad de la IA: Preventing harm and maintaining appropriate reliability and control.

AI security / protección de la IA frente a ataques: Protecting systems and assets against unauthorized access or malicious interference.

Governance / gobernanza: Institutions, rules, authority, and accountability shaping AI decisions.

General-purpose AI / IA de uso general: Systems usable across many tasks.

AGI / IAG: Hypothetical broadly capable AI; definitions and thresholds vary.

Superintelligence / superinteligencia: Capability substantially above humans across many relevant domains.

Alignment / alineación: Making behavior serve intended goals and constraints.

Agent / agente: A system that can pursue tasks through sequences of actions, often using tools.

Training / entrenamiento: Updating model parameters using data and computation.

Inference / inferencia: Running a trained model to produce outputs.

Compute / cómputo: Computational resources used in training or operation.

Benchmark / prueba de referencia: A defined set of tasks and scoring rules.

Evaluation / evaluación: A structured assessment of capabilities, behavior, or risks.

Hallucination / alucinación: Generated content that is false or unsupported; not necessarily deliberate.

Proxy / indicador sustituto: A measurable stand-in for a goal, which can diverge from it.

Specification gaming / aprovechamiento de fallas en la especificación: Meeting a stated metric or rule while missing the intended goal.

Misuse / uso malicioso: A person's deliberate harmful use of a system.

Threat model / modelo de amenazas: An account of actors, capabilities, access, and pathways to harm.

Causal pathway / vía causal: Linked steps explaining how an outcome could arise.

Loss of control / pérdida de control: A situation where humans cannot effectively direct or regain control of a system.

Defense in depth / defensa en profundidad: Multiple safeguards intended to prevent one failure from causing harm.

Correlated failure / fallo correlacionado: Failures that share a cause or dependency.

Least privilege / mínimo privilegio: Giving only the permissions needed for a bounded task.

Monitoring / monitoreo: Observing signals that may require action.

Fallback / alternativa de respaldo: A usable process when the primary system is unavailable or unsuitable.

Appeal / apelación: A route to challenge a decision and obtain review.

Forecast / predicción: A claim about the future, with assumptions and uncertainty.

Calibration / calibración: How stated confidence relates to actual correctness across many judgments.

Technical words are tools for clarity. Explain the mechanism before naming the term. “Safety” and “security” can both translate as “seguridad”; specify which meaning you intend.

Reading map and attribution

Core readings and offline equivalents

Module 1: International AI Safety Report summary, introduction and risk categories; alternate: module 1's three claim cards.

Module 2: METR limitations note; alternate: W2's evidence cards and evaluation design.

Module 3: International AI Safety Report summary, risk sections; alternate: module 3's classification cases and W3.

Module 4: International AI Safety Report summary, risk-management section; alternate: W4's safeguard cards.

Module 5: BlueDot contribution ideas; alternate: module 5's three project options.

Readings support 30-minute source labs, not an additional full course. If a link is unavailable, use the named in-package alternative. The official report summary is available in English and Spanish. Other external readings are primarily English; HASI's equivalent core tasks and explanations are fully bilingual.

Source links

BlueDot AGI Strategy curriculum

BlueDot independent reuse guidance, Joshua Landes, 2026

BlueDot shared AGI discussion template

BlueDot activity templates

BlueDot facilitator training

International AI Safety Report 2026: executive summary

METR: Clarifying limitations of time horizon, January 2026

Narayanan and Kapoor: AI as Normal Technology, April 2025

What HASI adapted

Retained: the five-stage progression, prepared discussion, peer explanation, causal risk analysis, layered defenses, and a personal contribution plan. Added: original Puente cases, explicit distinction between fact and scenario, college-level scaffolding, bilingual alternatives, assessment anchors, and host procedures. The shared BlueDot scenarios and their numerical future outcomes are not presented as forecasts or reused as factual evidence.

BlueDot's materials are supplied without warranty under its stated reuse conditions. Original third-party articles are linked, not reproduced or fully translated. The course hub may require a free account for some content; HASI's core activities do not depend on that account. Participant edition 1.1, September 20, 2026; sources checked September 12, 2026.