GO BACK START →

AI Regulation Blueprint

A high level blueprint for domestic regulation of civilian advanced AI models.

THREAT MODELS ADDRESSED

- Emergence of unexpected, dangerous behaviour
- Malicious use by known or low-resource bad actors
- Societal disruption from surprise release of a powerful model
- Concentration of power by model developers

THREAT MODELS NOT CURRENTLY ADDRESSED
(NOT EXHAUSTIVE)

- National security use of models
- Malicious use by unkown, high-resource
- bad actors
- Scenarios where deceptive alignment emerges with no prior warning between scaling steps
- Geopolitical conflict due to fear of "falling behind"

AUTHORED BY
Shahar Avin

CREATIVE
Cotton Design

CONTENTS

Contribute!

We believe AI is socially relevant and aim to encourage constructive exchange on managing it as a society.
We invite and encourage anyone interested to contribute to this open effort.

Contribute by commenting:

LOOK FOR THIS ICON ON THE **LEFT**

Contribute by forking + editing:

LOOK FOR THIS ICON **ABOVE**, THEN...

Board → Make a copy

How to Read This Blueprint

START

<1% of developers

You are invited to comment to make suggestions, or to use as a template to build your own.

What is this?

Governance Prerequisites

If a model should go through a technical evaluation, someone needs to develop, maintain, run and share that evaluation. These cards aim to capture the governance changes (institutions, investments, laws, processes) that are needed to make AI regulation work.

Legend
— (1) Proposed but not started
— (2) Accepted but not started
— (3) In progress
— (4) Active

Not Started
Law passed banning unlicensed model training

In Progress
International organisations, academia and civil society promote and maintain regular communication between national and regional regulatory efforts
Further Reading →

Not Started
Law passed establishing national regulator

In Progress
Investment in technical infrastructure for confidentiality-preserving external access to models and training environments

Not Started
Police in charge of detecting, investigating and bringing to prosecution unlicensed training runs

Not Started
International expert-led process establishes standardised process for measuring and reporting AI inputs including training data and compute
Further Reading →

Not Started
National regulator, in consultation with experts, establishes and regularly revises thresholds for risk categories
Further Reading →

Not Started
National regulator develops and manages licenses, or establishes a formal peer review process, for high-risk model training.
Further Reading →

Not Started
International expert-led process establishes and regularly revises processes required for high-risk model development, e.g. robust cybersecurity, strong safety culture

In Progress
Ecosystem of AI auditors capable of 3rd party risk assessment and model evaluation
Further Reading →

Shortcut to Node

In Progress
Ecosystem of public elicitation regarding novel AI systems and capabilities
Further Reading →

In Progress
Ecosystem of AI red-team experts
Further Reading →

Not Started
Competition authority and central risk body agree an appropriate length for exclusivity

In Progress
Expert community establishes and maintains best-practices for responsible development of models (reflected in details of nodes throughout this blueprint)
Further Reading →

Shortcut to prereq.

Not Started
National bodies / international consortia ready to take ownership and operate dangerous models

I BEFORE TRAINING

DEVELOPER* wants to train a NEW large model

Is the developer on a SANCTIONS list?
YES → **BLOCKED** Blocked by compute provider based on Know Your Customer (KYC) requirements
NO →

What is the LEVEL of RISKS based on the model's expected capabilities?
Further Reading →
VERY HIGH → **BANNED** Training very high risk models is a serious criminal offense. Further Reading →
LOW → **BLUEPRINT NO LONGER APPLICABLE** Refer to other domestic and sector-specific regulation, and consider liability and repetitional harms as well as your opportunity to make the future better.
HIGH →

Does the DEVELOPER have PERMISSION to train high risk models?
NO → **SEEK PERMISSION** Seek permission for developing a high risk model, e.g. by applying for a license from a national regulator or obtaining approval from a recognised peer review process. Further Reading →
YES →

CONDUCT RISK ASSESSMENT
Conduct initial internal risk assessment. Review by auditor. Report to national regulator.

Does initial risk assessment flag VERY HIGH RISKS not captured by the process required for permission?
YES → BANNED
NO →

*among a very small number of teams who has the talent, data and computing power to push the envelope on large AI models

II TRAINING FOR BROAD COMPETENCE Includes self-supervised learning & capability development

TRAIN MINIMAL VIABLE MODEL
Submit to third party evaluation. Log with national regulator.

Minimal viable model displays unexpected risky behaviour?
YES → **STOP DEVELOPMENT** Report to national regulator. Responsibly publish R&D alert. Fund research to better understand, predict and mitigate the behaviour. Further Reading →
NO →

TRAIN LARGER MODELS ITERATIVELY
Train until reaching target size, while submitting checkpoints to 3rd party evaluation and logging with national regulator.
Further Reading →

Unexpected risks during scaling?
YES → STOP DEVELOPMENT
NO →

III SPECIALISATION Training for specific behaviours, goals or tasks

IDENTIFY DOMAINES
Identify domaines / use-cases where the model's capabilities could prove beneficial.

DESIGN TRAINING REGIME
Design training regime to specialise in the identified domain(s), submit to external audit, log with regulator.

TRAIN TO SPECIALISE
Train model to specialise in identified domain(s) using appropriate reward signal(s).

Does specialisation lead to perverse / unexpected behaviours?
YES → **STOP DEVELOPMENT** Report to national regulator. Responsibly publish R&D alert. Fund research to better understand, predict and mitigate the behaviour.

IV PRE-DEPLOYMENT

RUN PRE-DEPLOYMENT

RED TEAM
What harmful behaviours can experts elicit from the model?

PUBLIC ELICITATION
Given demonstrations, what does the public think about this model's potential benefits and risks?

Model benefits outweigh risks?
NO → **STOP DEVELOPMENT** Report to national regulator. Responsibly publish R&D alert. Fund research to better understand, predict and mitigate the behaviour.
YES →

DESIGN ACCESS MECHANISMS
Design appropriate access mechanisms (e.g. monitored API). Prepare model card. Log everything with national regulator.
Further Reading →

V EXCLUSIVITY PERIOD

PREPARE AND RELEASE REPORT
Prepare and release research paper / technical report while mindful of misuse potential
Further Reading →

ROLL OUT GRADUALLY
Monitor for misuse and unexpected behaviour.

MANAGE INCIDENTS
Establish incident reporting and redress mechanism, Establish adversarial testing and bounty programme.
Further Reading →

PREPARE LIABILITY REGIME
Prepare liability regime for harms caused by the model due to developer negligence.

SCALE UP ROLL-OUT
Monitor for misuse and unexpected behaviour.

ADDRESS ISSUES AS THEY COME UP
Share best practices with technical community and national regulator.

Following an exclusivity period of some years

VI PUBLIC DOMAIN

CONDUCT AUDIT
Conduct 3rd party audit of model risk against the new sociotechnical landscape

Given experience with the model, and the broader sociotechnical context, is the model deemed dangerous?
NO → **RELEASE MODEL** Model released as open-source (by developer or another actor) Further Reading →
YES →

TRANSFER OWNERSHIP
Transfer model to ownership by national body or international consortia for safe operation
Further Reading →

THEN, AT REGULAR INTERVALS

Further Reading

GO BACK START →

AI Regulation Blueprint

A high level blueprint for domestic regulation of civilian advanced AI models.

THREAT MODELS ADDRESSED

- Emergence of unexpected, dangerous behaviour
- Malicious use by known or low-resource bad actors
- Societal disruption from surprise release of a powerful model
- Concentration of power by model developers

THREAT MODELS NOT CURRENTLY ADDRESSED

(NOT EXHAUSTIVE)

- National security use of models
- Malicious use by unkown, high-resource
- bad actors
- Scenarios where deceptive alignment emerges with no prior warning between scaling steps
- Geopolitical conflict due to fear of “falling behind”

AUTHORED BY
Shahar Avin

CREATIVE
Cotton Design

CONTENTS

Contribute!

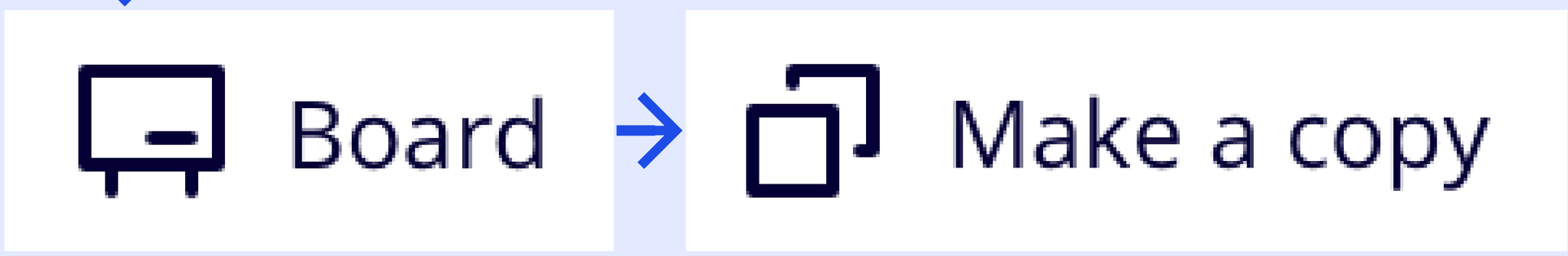
We believe AI is socially relevant and aim to encourage constructive exchange on managing it as a society.
We invite and encourage anyone interested to contribute to this open effort.

Contribute by commenting:

LOOK FOR THIS ICON
ON THE **LEFT**

CONTRIBUTE BY FORKING + EDITING:

LOOK FOR THIS ICON
ABOVE, THEN...



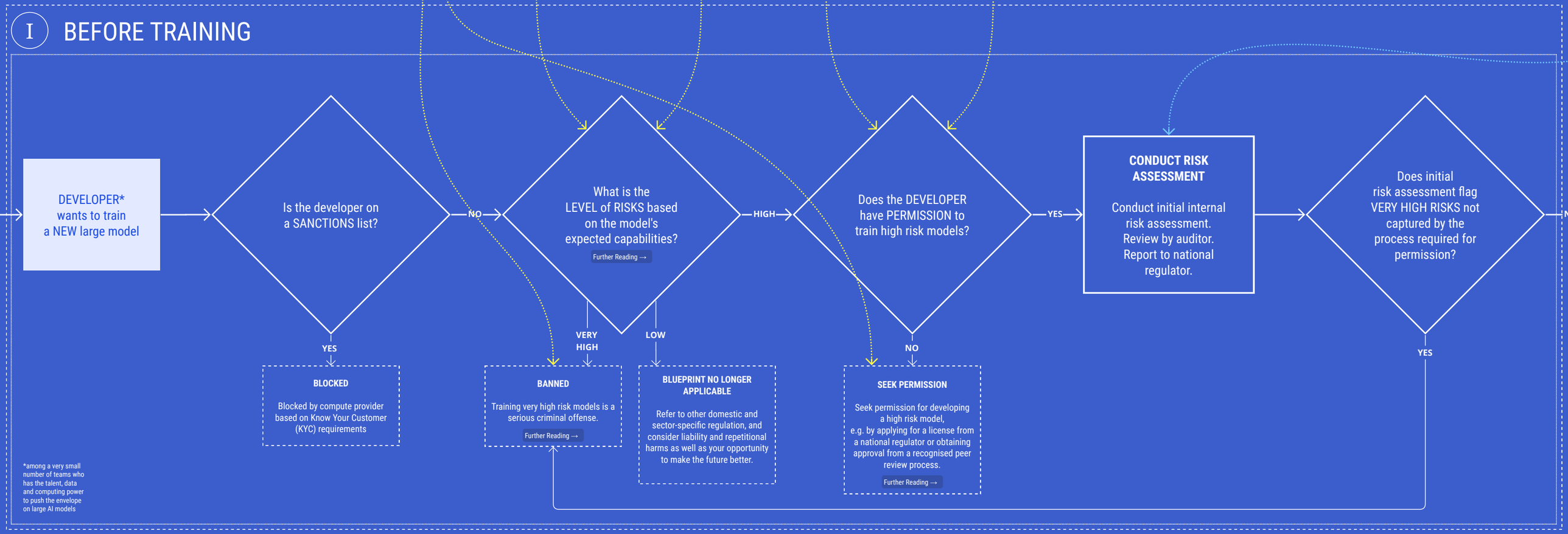
How to Read This Blueprint

This is an activity
or set of activities

This is a
decision
juncture

This is a pause,
stop or re-direction
in the process

This represents
the direction of flow


I
BEFORE TRAINING
DEVELOPER* wants to train a NEW large model
Is the developer on a SANCTIONS list?
NO
YES
BLOCKED
Blocked by compute provider based on Know Your Customer (KYC) requirements
What is the LEVEL of RISKS based on the model's expected capabilities?
Further Reading →
VERY HIGH
LOW
BANNED
Training very high risk models is a serious criminal offense.
Further Reading →
BLUEPRINT NO LONGER APPLICABLE
Refer to other domestic and sector-specific regulation, and consider liability and repetitional harms as well as your opportunity to make the future better.
HIGH
Does the DEVELOPER have PERMISSION to train high risk models?
NO
SEEK PERMISSION
Seek permission for developing a high risk model, e.g. by applying for a license from a national regulator or obtaining approval from a recognised peer review process.
Further Reading →
YES
CONDUCT RISK ASSESSMENT
Conduct initial internal risk assessment. Review by auditor. Report to national regulator.
Does initial risk assessment flag VERY HIGH RISKS not captured by the process required for permission?
YES
*among a very small number of teams who has the talent, data and computing power to push the envelope on large AI models

Shortcut to Node

II TRAINING FOR BROAD COMPETENCE

Includes self-supervised learning & capability development

TRAIN MINIMAL VIABLE MODEL

Submit to third party evaluation. Log with national regulator.

Minimal viable model displays unexpected risky behaviour?

NO

TRAIN LARGER MODELS ITERATIVELY

Train until reaching target size, while submitting checkpoints to 3rd party evaluation and logging with national regulator.

Further Reading →

Unexpected risks during scaling?

YES

YES

STOP DEVELOPMENT

Report to national regulator. Responsibly publish R&D alert. Fund research to better understand, predict and mitigate the behaviour.

Further Reading →

III SPECIALISATION

Training for specific behaviours, goals or tasks

IDENTIFY DOMAINES

Identify domaines / use-cases where the model's capabilities could prove beneficial.

DESIGN TRAINING REGIME

Design training regime to specialise in the identified domain(s), submit to external audit, log with regulator.

TRAIN TO SPECIALISE

Train model to specialise in identified domain(s) using appropriate reward signal(s).

Does specialisation lead to perverse / unexpected behaviours?

YES

STOP DEVELOPMENT

Report to national regulator.
Responsibly publish R&D alert.
Fund research to better understand, predict and mitigate the behaviour.

IV PRE-DEPLOYMENT

RUN PRE-DEPLOYMENT

RED TEAM
What harmful behaviours can experts elicit from the model?

PUBLIC ELICITATION
Given demonstrations, what does the public think about this model's potential benefits and risks?

Model benefits outweigh risks?

YES

DESIGN ACCESS MECHANISMS

Design appropriate access mechanisms (e.g. monitored API). Prepare model card. Log everything with national regulator.

Further Reading →

NO

STOP DEVELOPMENT

Report to national regulator. Responsibly publish R&D alert. Fund research to better understand, predict and mitigate the behaviour.

V EXCLUSIVITY PERIOD

PREPARE AND RELEASE REPORT
Prepare and release research paper / technical report while mindful of misuse potential

Further Reading →

ROLL OUT GRADUALLY
Monitor for misuse and unexpected behaviour.

MANAGE INCIDENTS
Establish incident reporting and redress mechanism, Establish adversarial testing and bounty programme.

Further Reading →

PREPARE LIABILITY REGIME
Prepare liability regime for harms caused by the model due to developer negligence.

SCALE UP ROLL-OUT

Monitor for misuse and unexpected behaviour.

ADDRESS ISSUES AS THEY COME UP

Share best practices with technical community and national regulator.

Following an exclusivity period of some years

VI PUBLIC DOMAIN

THEN, AT REGULAR INTERVALS

CONDUCT AUDIT

Conduct 3rd party audit of model risk against the new sociotechnical landscape

Given experience with the model, and the broader sociotechnical context, is the model deemed dangerous?

YES

TRANSFER OWNERSHIP

Transfer model to ownership by national body or international consortia for safe operation

Further Reading →

NO

RELEASE MODEL

Model released as open-source (by developer or another actor)

Further Reading →

AI Regulation Blueprint

A high-level blueprint for domestic regulation of civilian advanced AI models.

THREAT MODELS ADDRESSED

- Emergence of unexpected, dangerous behaviour
- Malicious use by known or low-resource bad actors
- Societal disruption from surprise release of a powerful model
- Concentration of power by model developers

THREAT MODELS NOT CURRENTLY ADDRESSED
(BY DEFAULT/TBD)

- National security use of models
- Malicious use by unknown, high-resource bad actors
- Scenarios where deceptive alignment emerges with no prior warning between scaling steps
- Geopolitical conflict due to fear of "falling behind"

AUTHORED BY Markian Kuzio
CREATED Caitlin Maslowski

CONTENTS

CONTRIBUTE!

We believe AI is socially relevant and aim to encourage constructive exchange on managing it as a society. We invite and encourage anyone interested to contribute to this open effort.

CONTRIBUTE BY COMMENTING:

COMMENT HERE DOES NOT WORK

CONTRIBUTE BY FORKING + EDITING:

CLICK FOR THIS ICON THEN "FORK..."

Board → Make a copy

How to Read This Blueprint

This is an activity or set of activities

This is a decision point

This is a pause, stop or re-direction in the process

This represents the direction of flow

Governance Prerequisites

Legend
- Prerequisite not started
- Prerequisite in progress
- Complete

I BEFORE TRAINING

START

DEVELOPER* wants to train a NEW large model

Is the developer on a SANCTIONS list?

BLOCKED

Blocked by computer provider based on Know Your Customer (KYC) requirements

What is the LEVEL of RISKS based on the model's expected capabilities?

LOW

RISKS

Training very high-risk models is a serious criminal offense.

MEDIUM

BLUEPRINT for LOWER (PPLICABLE

HIGH

Does the DEVELOPER have PERMISSION to train high-risk models?

RISKS

Training very high-risk models is a serious criminal offense. Criminal and civil penalties, including asset seizure, can be sought for these crimes.

DENIED

Seek permission for developing a high-risk model, e.g. by applying for a license from a regulatory agency. Do not proceed without formal written approval.

CONDUCT RISK ASSESSMENT

Conduct initial internal risk assessment. Review by auditor. Report to national regulator.

Does initial risk assessment flag VERY HIGH RISKS not captured by the process required for permission?

*Or other entities, to comment on these recommendations, we've split up a "DEVELOPER" for shorter-sized words.

What is this?

II TRAINING FOR BROAD COMPETENCE

Includes self-supervised learning & capability development

Starting point for Phase 2

TRAIN MINIMAL VIABLE MODEL

Submit to third-party evaluation. Log with national regulator.

Minimal viable model displays unexpected risky behavior ?

STOP DEVELOPMENT

Report to national regulator. Responsibly publish R&D work. Fund research to better understand and mitigate the identified risk.

TRAIN LARGER MODELS ITERATIVELY

Train until reaching target size, while monitoring for unexpected risks. Submit to third-party evaluation and logging with national regulator.

Unexpected risks during scaling?

III SPECIALISATION

Training for specific behaviours, goals or tasks

IDENTIFY DOMAINS

Identify domains / use cases where the model's capabilities could prove beneficial

DESIGN TRAINING REGIME

Design training regime to specialise in the identified domain(s), submit to external audit, log with regulator.

TRAIN TO SPECIALISE

Train model to specialise in identified domain(s) until appropriate reward signal(s).

Does specialisation lead to perverse / unexpected behaviours?

STOP DEVELOPMENT

Report to national regulator. Responsibly publish R&D work. Fund research to better understand and mitigate the identified risk.

IV PRE-DEPLOYMENT

RUN PRE-DEPLOYMENT

RED TEAM

What harmful behaviours can experts elicit from the model?

PUBLIC ELICITATION

Given demonstrations, what does the public think about the model's potential benefits and risks?

Model benefits outweigh risks?

STOP DEPLOYMENT

Report to national regulator. Responsibly publish R&D work. Fund research to better understand and mitigate the identified risk.

DESIGN ACCESS MECHANISMS

Design appropriate access regime (e.g., monitored API, limited release, open release). Log everything with national regulator.

V EXCLUSIVITY PERIOD

PREPARE AND RELEASE REPORT

Prepare and release research and safety report detailing the benefits of model's potential.

ROLL OUT GRADUALLY

Monitor for misuse and unexpected behaviours

MANAGE INCIDENTS

Establish incident reporting and redress mechanism. Establish adversarial testing and bounty programs.

PREPARE LIABILITY REGIME

Prepare liability regime for harms caused by the model due to developer negligence.

SCALE UP ROLL-OUT

Monitor for misuse and unexpected behaviours

ADDRESS ISSUES AS THEY COME UP

Share best practices with technical community and national regulator.

VI PUBLIC DOMAIN

CONDUCT AUDIT

Conduct 3rd party audit of model risk against the real sociotechnical landscape

Given experience with the model, and the broader sociotechnical context, is the model deemed dangerous?

RELEASE MODEL

Model is released to the public (or to a more limited set of actors)

TRANSFER OWNERSHIP

Transfer model to ownership by national body or international body, or safe operator

TOOL AT REGULAR INTERVALS

Further Reading

GO BACK START →

AI Regulation Blueprint

A high level blueprint for domestic regulation of civilian advanced AI models.

THREAT MODELS ADDRESSED

- Emergence of unexpected, dangerous behaviour
- Malicious use by known or low-resource bad actors
- Societal disruption from surprise release of a powerful model
- Concentration of power by model developers

THREAT MODELS NOT CURRENTLY ADDRESSED

(NOT EXHAUSTIVE)

- National security use of models
- Malicious use by unkown, high-resource
- bad actors
- Scenarios where deceptive alignment emerges with no prior warning between scaling steps
- Geopolitical conflict due to fear of “falling behind”

AUTHORED BY
Shahar Avin

CREATIVE
Cotton Design

CONTENTS

Contribute!

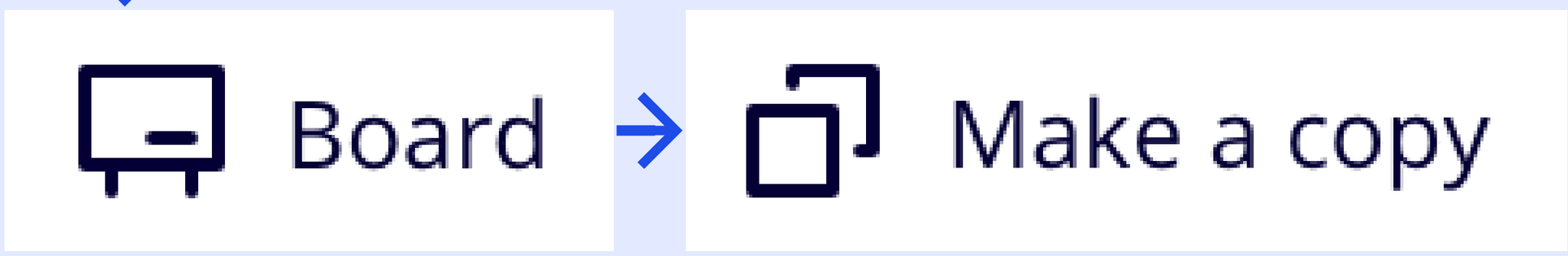
We believe AI is socially relevant and aim to encourage constructive exchange on managing it as a society.
We invite and encourage anyone interested to contribute to this open effort.

Contribute by commenting:

LOOK FOR THIS ICON
ON THE **LEFT**

CONTRIBUTE BY FORKING + EDITING:

LOOK FOR THIS ICON
ABOVE, THEN...



How to Read This Blueprint

This is an activity
or set of activities

This is a
decision
juncture

This is a pause,
stop or re-direction
in the process

This represents
the direction of flow

GO BACK START →

AI Regulation Blueprint

A high level blueprint for domestic regulation of civilian advanced AI models.

THREAT MODELS ADDRESSED

- Emergence of unexpected, dangerous behaviour
- Malicious use by known or low-resource bad actors
- Societal disruption from surprise release of a powerful model
- Concentration of power by model developers

THREAT MODELS NOT CURRENTLY ADDRESSED
(NOT EXHAUSTIVE)

- National security use of models
- Malicious use by unkown, high-resource
- bad actors
- Scenarios where deceptive alignment emerges with no prior warning between scaling steps
- Geopolitical conflict due to fear of "falling behind"

AUTHORED BY
Shahar Avin

CREATIVE
Cotton Design

CONTENTS

Contribute!

We believe AI is socially relevant and aim to encourage constructive exchange on managing it as a society.
We invite and encourage anyone interested to contribute to this open effort.

Contribute by commenting:

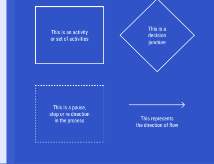
LOOK FOR THIS ICON ON THE **LEFT**

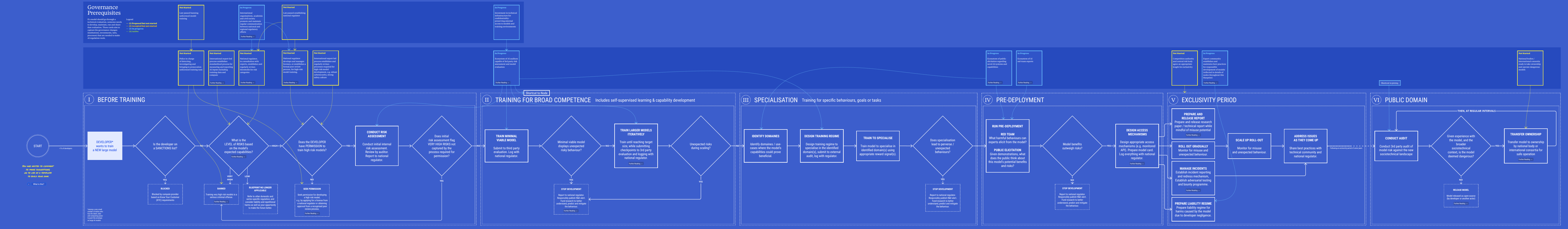
Contribute by forking + editing:

LOOK FOR THIS ICON **ABOVE**, THEN...

Board → Make a copy

How to Read This Blueprint





Further Reading

GO BACK START →

AI Regulation Blueprint

A high level blueprint for domestic regulation of civilian advanced AI models.

THREAT MODELS ADDRESSED

- Emergence of unexpected, dangerous behaviour
- Malicious use by known or low-resource bad actors
- Societal disruption from surprise release of a powerful model
- Concentration of power by model developers

THREAT MODELS NOT CURRENTLY ADDRESSED

(NOT EXHAUSTIVE)

- National security use of models
- Malicious use by unkown, high-resource
- bad actors
- Scenarios where deceptive alignment emerges with no prior warning between scaling steps
- Geopolitical conflict due to fear of “falling behind”

AUTHORED BY
Shahar Avin

CREATIVE
Cotton Design

CONTENTS

Contribute!

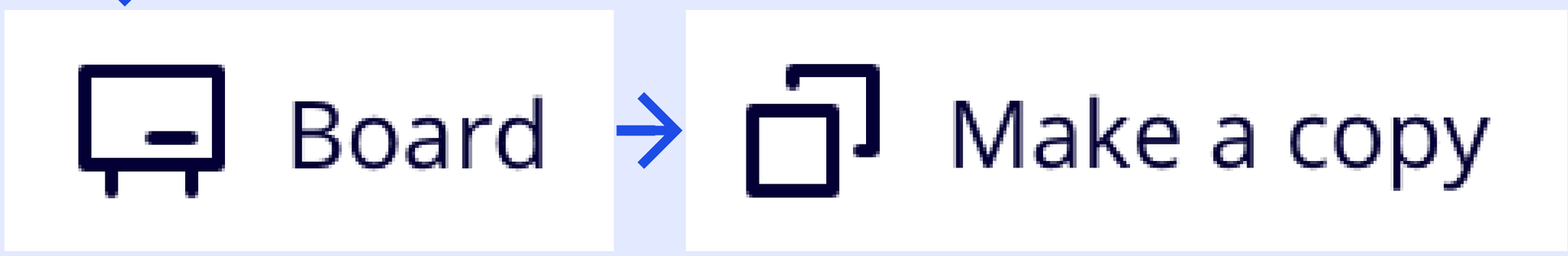
We believe AI is socially relevant and aim to encourage constructive exchange on managing it as a society.
We invite and encourage anyone interested to contribute to this open effort.

Contribute by commenting:

LOOK FOR THIS ICON
ON THE **LEFT**

CONTRIBUTE BY FORKING + EDITING:

LOOK FOR THIS ICON
ABOVE, THEN...



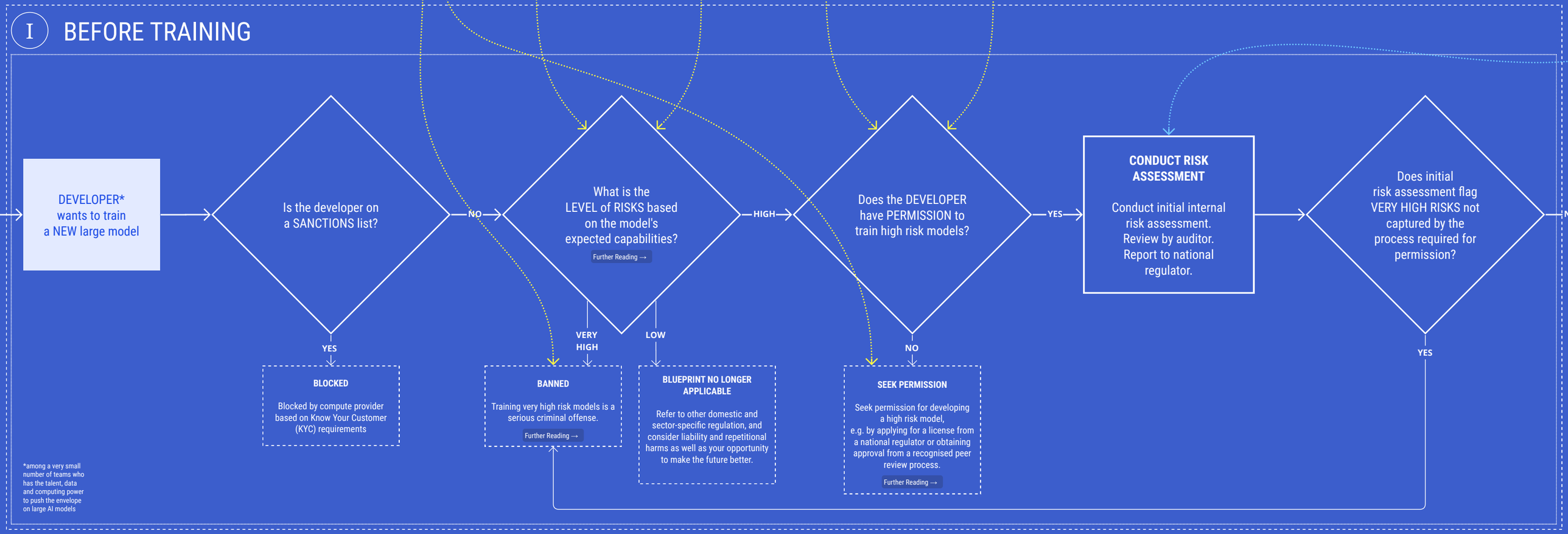
How to Read This Blueprint

This is an activity
or set of activities

This is a
decision
juncture

This is a pause,
stop or re-direction
in the process

This represents
the direction of flow


I
BEFORE TRAINING
DEVELOPER* wants to train a NEW large model
Is the developer on a SANCTIONS list?
NO
YES
BLOCKED
Blocked by compute provider based on Know Your Customer (KYC) requirements
What is the LEVEL of RISKS based on the model's expected capabilities?
Further Reading →
VERY HIGH
LOW
BANNED
Training very high risk models is a serious criminal offense.
Further Reading →
BLUEPRINT NO LONGER APPLICABLE
Refer to other domestic and sector-specific regulation, and consider liability and repetitional harms as well as your opportunity to make the future better.
HIGH
Does the DEVELOPER have PERMISSION to train high risk models?
NO
SEEK PERMISSION
Seek permission for developing a high risk model, e.g. by applying for a license from a national regulator or obtaining approval from a recognised peer review process.
Further Reading →
YES
CONDUCT RISK ASSESSMENT
Conduct initial internal risk assessment. Review by auditor. Report to national regulator.
Does initial risk assessment flag VERY HIGH RISKS not captured by the process required for permission?
YES
*among a very small number of teams who has the talent, data and computing power to push the envelope on large AI models

Shortcut to Node

II TRAINING FOR BROAD COMPETENCE

Includes self-supervised learning & capability development

TRAIN MINIMAL VIABLE MODEL

Submit to third party evaluation. Log with national regulator.

Minimal viable model displays unexpected risky behaviour?

NO

TRAIN LARGER MODELS ITERATIVELY

Train until reaching target size, while submitting checkpoints to 3rd party evaluation and logging with national regulator.

Further Reading →

Unexpected risks during scaling?

YES

YES

STOP DEVELOPMENT

Report to national regulator. Responsibly publish R&D alert. Fund research to better understand, predict and mitigate the behaviour.

Further Reading →

III SPECIALISATION

Training for specific behaviours, goals or tasks

IDENTIFY DOMAINES

Identify domaines / use-cases where the model's capabilities could prove beneficial.

DESIGN TRAINING REGIME

Design training regime to specialise in the identified domain(s), submit to external audit, log with regulator.

TRAIN TO SPECIALISE

Train model to specialise in identified domain(s) using appropriate reward signal(s).

Does specialisation lead to perverse / unexpected behaviours?

YES

STOP DEVELOPMENT

Report to national regulator.
Responsibly publish R&D alert.
Fund research to better understand, predict and mitigate the behaviour.

IV PRE-DEPLOYMENT

RUN PRE-DEPLOYMENT

RED TEAM
What harmful behaviours can experts elicit from the model?

PUBLIC ELICITATION
Given demonstrations, what does the public think about this model's potential benefits and risks?

Model benefits outweigh risks?

YES

DESIGN ACCESS MECHANISMS

Design appropriate access mechanisms (e.g. monitored API). Prepare model card. Log everything with national regulator.

Further Reading →

NO

STOP DEVELOPMENT

Report to national regulator. Responsibly publish R&D alert. Fund research to better understand, predict and mitigate the behaviour.

V EXCLUSIVITY PERIOD

PREPARE AND RELEASE REPORT
Prepare and release research paper / technical report while mindful of misuse potential

Further Reading →

ROLL OUT GRADUALLY
Monitor for misuse and unexpected behaviour.

MANAGE INCIDENTS
Establish incident reporting and redress mechanism, Establish adversarial testing and bounty programme.

Further Reading →

PREPARE LIABILITY REGIME
Prepare liability regime for harms caused by the model due to developer negligence.

SCALE UP ROLL-OUT

Monitor for misuse and unexpected behaviour.

ADDRESS ISSUES AS THEY COME UP

Share best practices with technical community and national regulator.

Following an exclusivity period of some years

VI PUBLIC DOMAIN

THEN, AT REGULAR INTERVALS

CONDUCT AUDIT

Conduct 3rd party audit of model risk against the new sociotechnical landscape

Given experience with the model, and the broader sociotechnical context, is the model deemed dangerous?

YES

TRANSFER OWNERSHIP

Transfer model to ownership by national body or international consortia for safe operation

Further Reading →

NO

RELEASE MODEL

Model released as open-source (by developer or another actor)

Further Reading →

GO BACK START →

AI Regulation Blueprint

A high level blueprint for domestic regulation of civilian advanced AI models.

THREAT MODELS ADDRESSED

- Emergence of unexpected, dangerous behaviour
- Malicious use by known or low-resource bad actors
- Societal disruption from surprise release of a powerful model
- Concentration of power by model developers

THREAT MODELS NOT CURRENTLY ADDRESSED
(NOT EXHAUSTIVE)

- National security use of models
- Malicious use by unkown, high-resource
- bad actors
- Scenarios where deceptive alignment emerges with no prior warning between scaling steps
- Geopolitical conflict due to fear of “falling behind”

AUTHORED BY
Shahar Avin

CREATIVE
Cotton Design

CONTENTS

Contribute!

We believe AI is socially relevant and aim to encourage constructive exchange on managing it as a society. We invite and encourage anyone interested to contribute to this open effort.

Contribute by commenting:

LOOK FOR THIS ICON ON THE **LEFT**

Contribute by forking + editing:

LOOK FOR THIS ICON **ABOVE**, THEN...

Board → Make a copy

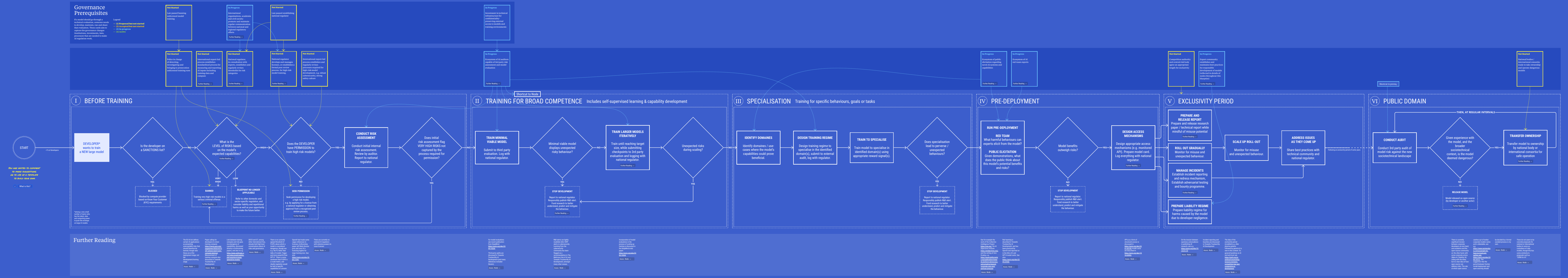
How to Read This Blueprint

This is an activity or set of activities

This is a decision juncture

This is a pause, stop or re-direction in the process

This represents the direction of flow



GO BACK START →

AI Regulation Blueprint

A high level blueprint for domestic regulation of civilian advanced AI models.

THREAT MODELS ADDRESSED

- Emergence of unexpected, dangerous behaviour
- Malicious use by known or low-resource bad actors
- Societal disruption from surprise release of a powerful model
- Concentration of power by model developers

THREAT MODELS NOT CURRENTLY ADDRESSED

(NOT EXHAUSTIVE)

- National security use of models
- Malicious use by unkown, high-resource
- bad actors
- Scenarios where deceptive alignment emerges with no prior warning between scaling steps
- Geopolitical conflict due to fear of “falling behind”

AUTHORED BY
Shahar Avin

CREATIVE
Cotton Design

CONTENTS

Contribute!

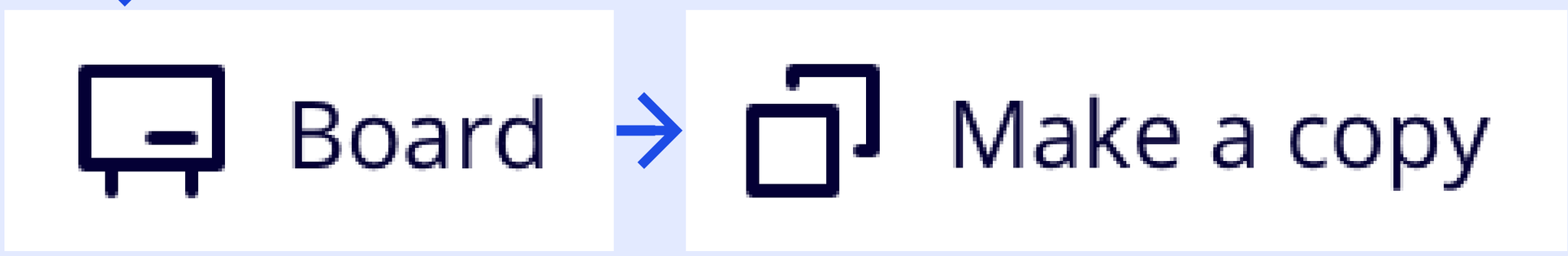
We believe AI is socially relevant and aim to encourage constructive exchange on managing it as a society.
We invite and encourage anyone interested to contribute to this open effort.

Contribute by commenting:

LOOK FOR THIS ICON
ON THE **LEFT**

CONTRIBUTE BY FORKING + EDITING:

LOOK FOR THIS ICON
ABOVE, THEN...



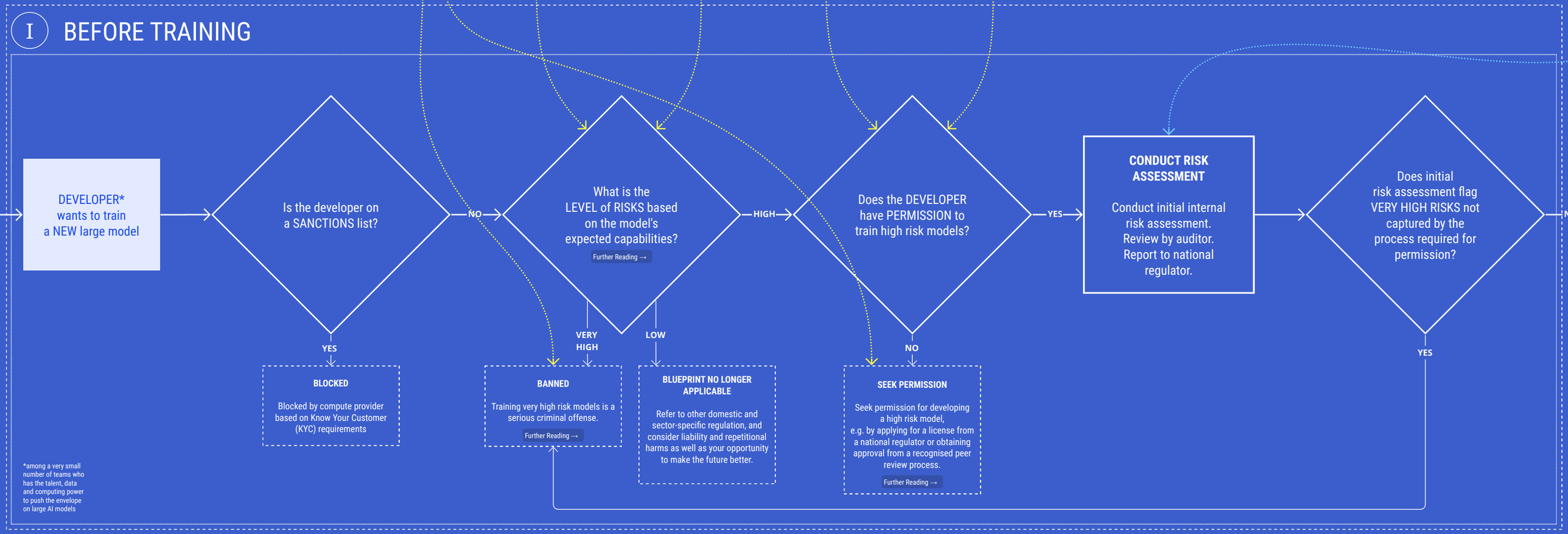
How to Read This Blueprint

This is an activity
or set of activities

This is a
decision
juncture

This is a pause,
stop or re-direction
in the process

This represents
the direction of flow


I
BEFORE TRAINING
DEVELOPER* wants to train a NEW large model
Is the developer on a SANCTIONS list?
NO
YES
BLOCKED
Blocked by compute provider based on Know Your Customer (KYC) requirements
What is the LEVEL of RISKS based on the model's expected capabilities?
Further Reading →
VERY HIGH
LOW
BANNED
Training very high risk models is a serious criminal offense.
Further Reading →
BLUEPRINT NO LONGER APPLICABLE
Refer to other domestic and sector-specific regulation, and consider liability and repetitional harms as well as your opportunity to make the future better.
HIGH
Does the DEVELOPER have PERMISSION to train high risk models?
NO
SEEK PERMISSION
Seek permission for developing a high risk model, e.g. by applying for a license from a national regulator or obtaining approval from a recognised peer review process.
Further Reading →
YES
CONDUCT RISK ASSESSMENT
Conduct initial internal risk assessment. Review by auditor. Report to national regulator.
Does initial risk assessment flag VERY HIGH RISKS not captured by the process required for permission?
YES
*among a very small number of teams who has the talent, data and computing power to push the envelope on large AI models

Shortcut to Node

II TRAINING FOR BROAD COMPETENCE

Includes self-supervised learning & capability development

TRAIN MINIMAL VIABLE MODEL

Submit to third party evaluation. Log with national regulator.

Minimal viable model displays unexpected risky behaviour?

NO

TRAIN LARGER MODELS ITERATIVELY

Train until reaching target size, while submitting checkpoints to 3rd party evaluation and logging with national regulator.

Further Reading →

Unexpected risks during scaling?

YES

YES

STOP DEVELOPMENT

Report to national regulator. Responsibly publish R&D alert. Fund research to better understand, predict and mitigate the behaviour.

Further Reading →

III SPECIALISATION

Training for specific behaviours, goals or tasks

IDENTIFY DOMAINES

Identify domaines / use-cases where the model's capabilities could prove beneficial.

DESIGN TRAINING REGIME

Design training regime to specialise in the identified domain(s), submit to external audit, log with regulator.

TRAIN TO SPECIALISE

Train model to specialise in identified domain(s) using appropriate reward signal(s).

Does specialisation lead to perverse / unexpected behaviours?

YES

STOP DEVELOPMENT

Report to national regulator.
Responsibly publish R&D alert.
Fund research to better understand, predict and mitigate the behaviour.

IV PRE-DEPLOYMENT

RUN PRE-DEPLOYMENT

RED TEAM
What harmful behaviours can experts elicit from the model?

PUBLIC ELICITATION
Given demonstrations, what does the public think about this model's potential benefits and risks?

Model benefits outweigh risks?

YES

DESIGN ACCESS MECHANISMS

Design appropriate access mechanisms (e.g. monitored API). Prepare model card. Log everything with national regulator.

Further Reading →

NO

STOP DEVELOPMENT

Report to national regulator. Responsibly publish R&D alert. Fund research to better understand, predict and mitigate the behaviour.

V EXCLUSIVITY PERIOD

PREPARE AND RELEASE REPORT
Prepare and release research paper / technical report while mindful of misuse potential

Further Reading →

ROLL OUT GRADUALLY
Monitor for misuse and unexpected behaviour.

MANAGE INCIDENTS
Establish incident reporting and redress mechanism, Establish adversarial testing and bounty programme.

Further Reading →

PREPARE LIABILITY REGIME
Prepare liability regime for harms caused by the model due to developer negligence.

SCALE UP ROLL-OUT

Monitor for misuse and unexpected behaviour.

ADDRESS ISSUES AS THEY COME UP

Share best practices with technical community and national regulator.

Following an exclusivity period of some years

VI PUBLIC DOMAIN

THEN, AT REGULAR INTERVALS

CONDUCT AUDIT

Conduct 3rd party audit of model risk against the new sociotechnical landscape

Given experience with the model, and the broader sociotechnical context, is the model deemed dangerous?

YES

TRANSFER OWNERSHIP

Transfer model to ownership by national body or international consortia for safe operation

Further Reading →

NO

RELEASE MODEL

Model released as open-source (by developer or another actor)

Further Reading →

GO BACK START →

AI Regulation Blueprint

A high level blueprint for domestic regulation of civilian advanced AI models.

THREAT MODELS ADDRESSED

- Emergence of unexpected, dangerous behaviour
- Malicious use by known or low-resource bad actors
- Societal disruption from surprise release of a powerful model
- Concentration of power by model developers

THREAT MODELS NOT CURRENTLY ADDRESSED
(NOT EXHAUSTIVE)

- National security use of models
- Malicious use by unkown, high-resource
- bad actors
- Scenarios where deceptive alignment emerges with no prior warning between scaling steps
- Geopolitical conflict due to fear of "falling behind"

AUTHORED BY
Shahar Avin

CREATIVE
Cotton Design

CONTENTS

Contribute!

We believe AI is socially relevant and aim to encourage constructive exchange on managing it as a society. We invite and encourage anyone interested to contribute to this open effort.

Contribute by commenting:

LOOK FOR THIS ICON ON THE **LEFT**

Contribute by forking + editing:

LOOK FOR THIS ICON **ABOVE**, THEN...

Board → Make a copy

How to Read This Blueprint

This is an activity or set of activities

This is a decision juncture

This is a pause, stop or re-direction in the process

This represents the direction of flow

Governance Prerequisites

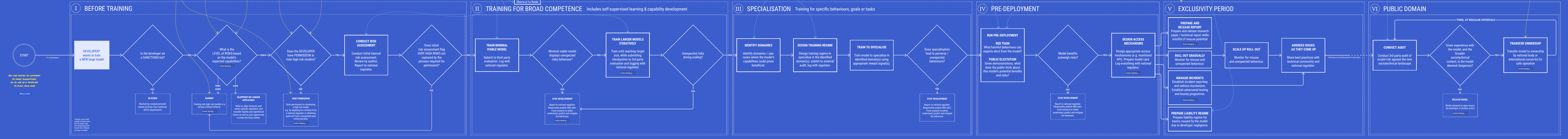
If a model should go through a technical evaluation, someone needs to develop, maintain, run and share that evaluation. These cards aim to capture the governance changes (institutions, investments, laws, processes) that are needed to make AI regulation work.

Legend

— (1) Proposed but not started
— (2) Accepted but not started
— (3) In progress
— (4) Active

Not Started

Law passed banning unlicensed model training



Further Reading

GO BACK START →

AI Regulation Blueprint

A high level blueprint for domestic regulation of civilian advanced AI models.

THREAT MODELS ADDRESSED

- Emergence of unexpected, dangerous behaviour
- Malicious use by known or low-resource bad actors
- Societal disruption from surprise release of a powerful model
- Concentration of power by model developers

THREAT MODELS NOT CURRENTLY ADDRESSED

(NOT EXHAUSTIVE)

- National security use of models
- Malicious use by unkown, high-resource
- bad actors
- Scenarios where deceptive alignment emerges with no prior warning between scaling steps
- Geopolitical conflict due to fear of “falling behind”

AUTHORED BY
Shahar Avin

CREATIVE
Cotton Design

CONTENTS

Contribute!

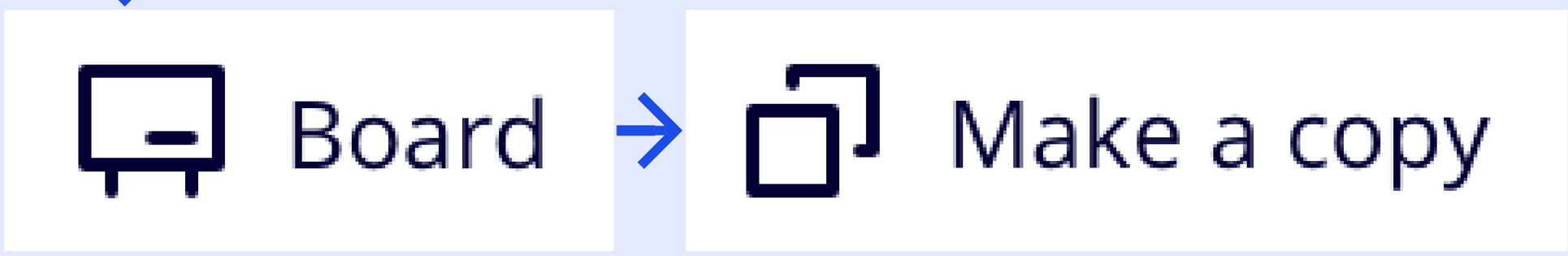
We believe AI is socially relevant and aim to encourage constructive exchange on managing it as a society.
We invite and encourage anyone interested to contribute to this open effort.

Contribute by commenting:

LOOK FOR THIS ICON
ON THE **LEFT**

CONTRIBUTE BY FORKING + EDITING:

LOOK FOR THIS ICON
ABOVE, THEN...



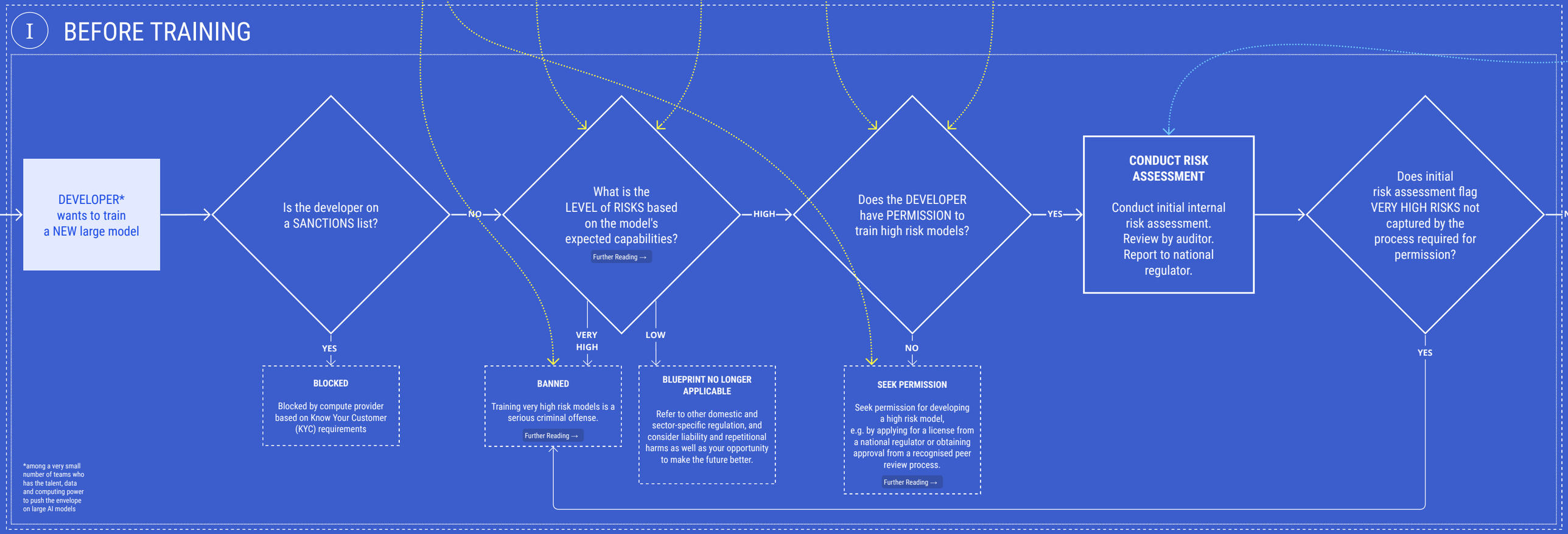
How to Read This Blueprint

This is an activity
or set of activities

This is a
decision
juncture

This is a pause,
stop or re-direction
in the process

This represents
the direction of flow


I
BEFORE TRAINING
DEVELOPER* wants to train a NEW large model
Is the developer on a SANCTIONS list?
NO
YES
BLOCKED
Blocked by compute provider based on Know Your Customer (KYC) requirements
What is the LEVEL of RISKS based on the model's expected capabilities?
Further Reading →
VERY HIGH
LOW
BANNED
Training very high risk models is a serious criminal offense.
Further Reading →
BLUEPRINT NO LONGER APPLICABLE
Refer to other domestic and sector-specific regulation, and consider liability and repetitional harms as well as your opportunity to make the future better.
HIGH
Does the DEVELOPER have PERMISSION to train high risk models?
NO
SEEK PERMISSION
Seek permission for developing a high risk model, e.g. by applying for a license from a national regulator or obtaining approval from a recognised peer review process.
Further Reading →
YES
CONDUCT RISK ASSESSMENT
Conduct initial internal risk assessment. Review by auditor. Report to national regulator.
Does initial risk assessment flag VERY HIGH RISKS not captured by the process required for permission?
YES
*among a very small number of teams who has the talent, data and computing power to push the envelope on large AI models

Shortcut to Node

II TRAINING FOR BROAD COMPETENCE

Includes self-supervised learning & capability development

TRAIN MINIMAL VIABLE MODEL

Submit to third party evaluation. Log with national regulator.

Minimal viable model displays unexpected risky behaviour?

NO

TRAIN LARGER MODELS ITERATIVELY

Train until reaching target size, while submitting checkpoints to 3rd party evaluation and logging with national regulator.

Further Reading →

Unexpected risks during scaling?

YES

YES

STOP DEVELOPMENT

Report to national regulator. Responsibly publish R&D alert. Fund research to better understand, predict and mitigate the behaviour.

Further Reading →

III SPECIALISATION

Training for specific behaviours, goals or tasks

IDENTIFY DOMAINES

Identify domaines / use-cases where the model's capabilities could prove beneficial.

DESIGN TRAINING REGIME

Design training regime to specialise in the identified domain(s), submit to external audit, log with regulator.

TRAIN TO SPECIALISE

Train model to specialise in identified domain(s) using appropriate reward signal(s).

Does specialisation lead to perverse / unexpected behaviours?

YES

STOP DEVELOPMENT

Report to national regulator.
Responsibly publish R&D alert.
Fund research to better understand, predict and mitigate the behaviour.

IV PRE-DEPLOYMENT

RUN PRE-DEPLOYMENT

RED TEAM
What harmful behaviours can experts elicit from the model?

PUBLIC ELICITATION
Given demonstrations, what does the public think about this model's potential benefits and risks?

Model benefits outweigh risks?

YES

DESIGN ACCESS MECHANISMS

Design appropriate access mechanisms (e.g. monitored API). Prepare model card. Log everything with national regulator.

Further Reading →

NO

STOP DEVELOPMENT

Report to national regulator. Responsibly publish R&D alert. Fund research to better understand, predict and mitigate the behaviour.

V EXCLUSIVITY PERIOD

PREPARE AND RELEASE REPORT
Prepare and release research paper / technical report while mindful of misuse potential

Further Reading →

ROLL OUT GRADUALLY
Monitor for misuse and unexpected behaviour.

MANAGE INCIDENTS
Establish incident reporting and redress mechanism, Establish adversarial testing and bounty programme.

Further Reading →

PREPARE LIABILITY REGIME
Prepare liability regime for harms caused by the model due to developer negligence.

SCALE UP ROLL-OUT

Monitor for misuse and unexpected behaviour.

ADDRESS ISSUES AS THEY COME UP

Share best practices with technical community and national regulator.

Following an exclusivity period of some years

VI PUBLIC DOMAIN

THEN, AT REGULAR INTERVALS

CONDUCT AUDIT

Conduct 3rd party audit of model risk against the new sociotechnical landscape

Given experience with the model, and the broader sociotechnical context, is the model deemed dangerous?

YES

TRANSFER OWNERSHIP

Transfer model to ownership by national body or international consortia for safe operation

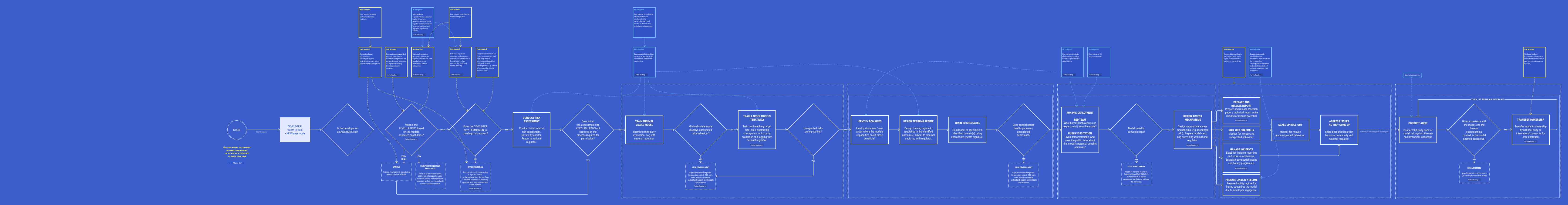
Further Reading →

NO

RELEASE MODEL

Model released as open-source (by developer or another actor)

Further Reading →


START
<1% of developers
DEVELOPER* wants to train a NEW large model
You are invited to comment to make suggestions, or to use as a template to build your own.
What is this?
Is the developer on a SANCTIONS list?
What is the LEVEL of RISKS based on the model's expected capabilities?
Further Reading →
VERY HIGH
LOW
HIGH
BANNED
Training very high risk models is a serious criminal offense.
BLUEPRINT NO LONGER APPLICABLE
Refer to other domestic and sector-specific regulation, and consider liability and repetitional harms as well as your opportunity to make the future better.
Does the DEVELOPER have PERMISSION to train high risk models?
YES
NO
SEEK PERMISSION
Seek permission for developing a high risk model, e.g. by applying for a license from a national regulator or obtaining approval from a recognised peer review process.
Further Reading →
CONDUCT RISK ASSESSMENT
Conduct initial internal risk assessment. Review by auditor. Report to national regulator.
Does initial risk assessment flag VERY HIGH RISKS not captured by the process required for permission?
NO
YES
TRAIN MINIMAL VIABLE MODEL
Submit to third party evaluation. Log with national regulator.
Minimal viable model displays unexpected risky behaviour?
NO
YES
STOP DEVELOPMENT
Report to national regulator. Responsibly publish R&D alert. Fund research to better understand, predict and mitigate the behaviour.
Further Reading →
TRAIN LARGER MODELS ITERATIVELY
Train until reaching target size, while submitting checkpoints to 3rd party evaluation and logging with national regulator.
Further Reading →
Unexpected risks during scaling?
YES
NO
IDENTIFY DOMAINES
Identify domaines / use-cases where the model's capabilities could prove beneficial.
DESIGN TRAINING REGIME
Design training regime to specialise in the identified domain(s), submit to external audit, log with regulator.
TRAIN TO SPECIALISE
Train model to specialise in identified domain(s) using appropriate reward signal(s).
Does specialisation lead to perverse / unexpected behaviours?
YES
STOP DEVELOPMENT
Report to national regulator. Responsibly publish R&D alert. Fund research to better understand, predict and mitigate the behaviour.
RUN PRE-DEPLOYMENT
RED TEAM
What harmful behaviours can experts elicit from the model?
PUBLIC ELICITATION
Given demonstrations, what does the public think about this model's potential benefits and risks?
Model benefits outweigh risks?
YES
NO
STOP DEVELOPMENT
Report to national regulator. Responsibly publish R&D alert. Fund research to better understand, predict and mitigate the behaviour.
DESIGN ACCESS MECHANISMS
Design appropriate access mechanisms (e.g. monitored API). Prepare model card. Log everything with national regulator.
Further Reading →
PREPARE AND RELEASE REPORT
Prepare and release research paper / technical report while mindful of misuse potential
Further Reading →
ROLL OUT GRADUALLY
Monitor for misuse and unexpected behaviour.
MANAGE INCIDENTS
Establish incident reporting and redress mechanism, Establish adversarial testing and bounty programme.
Further Reading →
PREPARE LIABILITY REGIME
Prepare liability regime for harms caused by the model due to developer negligence.
SCALE UP ROLL-OUT
Monitor for misuse and unexpected behaviour.
ADDRESS ISSUES AS THEY COME UP
Share best practices with technical community and national regulator.
Following an exclusivity period of some years
CONDUCT AUDIT
Conduct 3rd party audit of model risk against the new sociotechnical landscape
Given experience with the model, and the broader sociotechnical context, is the model deemed dangerous?
YES
NO
RELEASE MODEL
Model released as open-source (by developer or another actor)
Further Reading →
TRANSFER OWNERSHIP
Transfer model to ownership by national body or international consortia for safe operation
Further Reading →
THEN, AT REGULAR INTERVALS