The Elements of Decision Alignment

Mark S. Miller
Bill Tulloh

UCI 2017

The Elements of Decision Alignment

Mark S. Miller (**cs**)
Bill Tulloh (**econ**)

UCI 2017

When one object makes a request of another object, why do we expect the second object’s behavior to satisfy the first object’s wishes?

Networks of entities making requests of other entities:

- Object-oriented programs
- Human organizations
- Human economies

Borrowing Ideas from Economics

“Like an economist … we are interested in individual agents not so much for what they are internally as for what they have to offer each other

… much of object-oriented design is indeed *Design by Contract.”*

Object-Oriented Software Construction
Bertrand Meyer, 1997

Overview

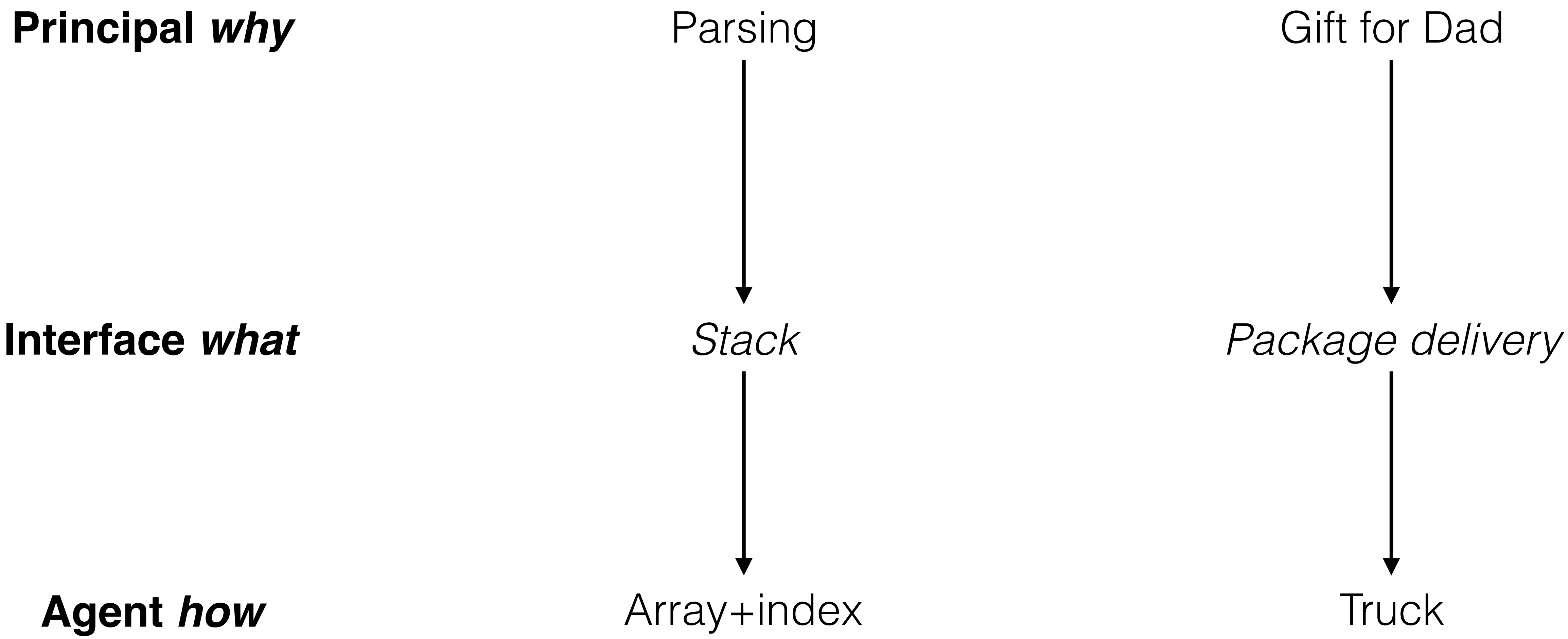
Making Requests

Aligning Decisions

Making Tradeoffs

Dividing and Composing Knowledge

Making Requests

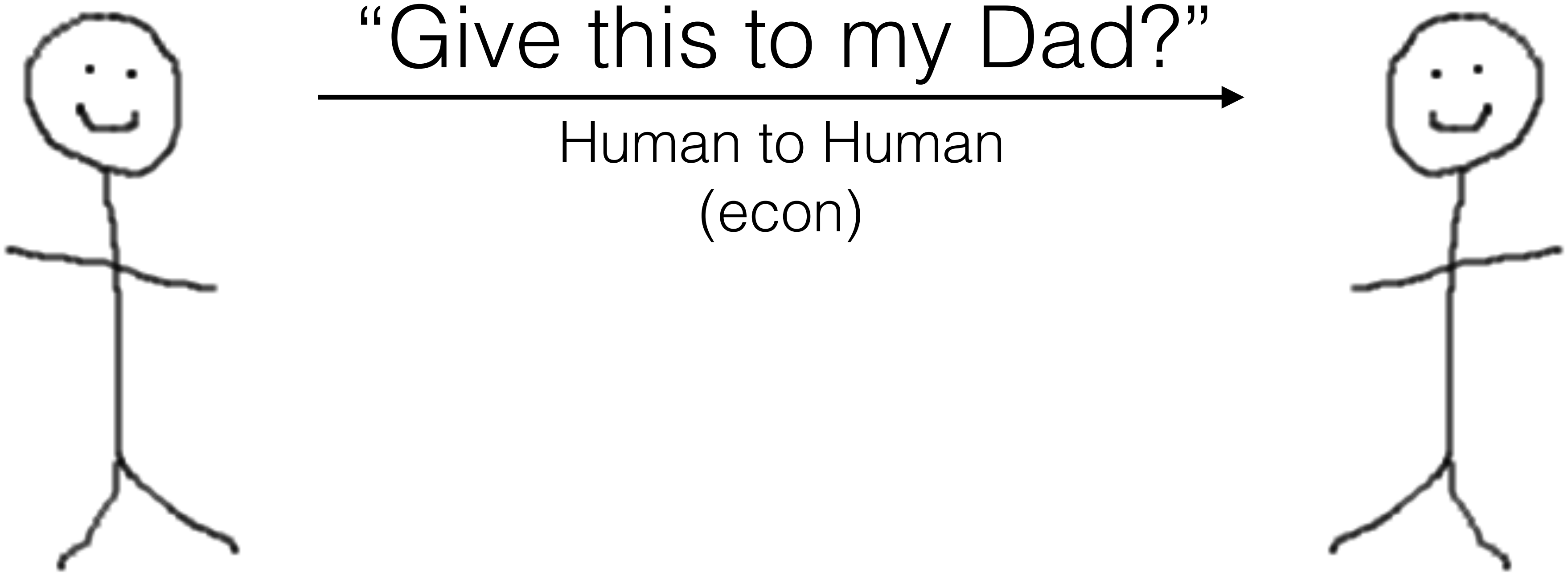


Principal *why* Parsing Gift for Dad

Interface *what* *Stack* *Package delivery*

Agent *how* Array+index Truck

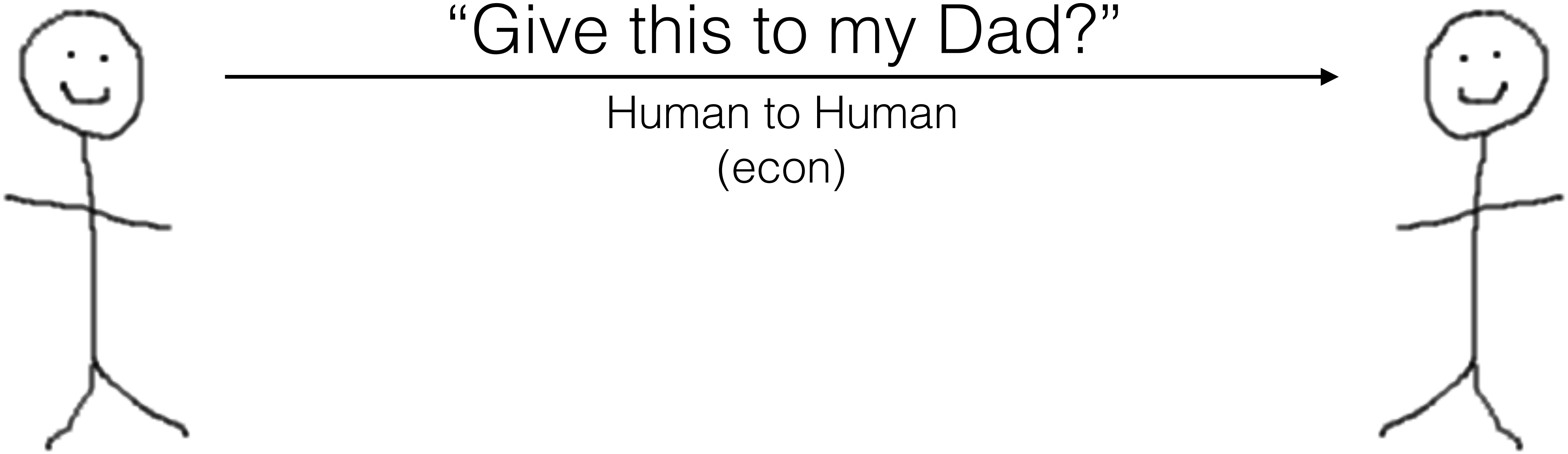
Principal-Agent relationships



“Give this to my Dad?”

Human to Human
(econ)

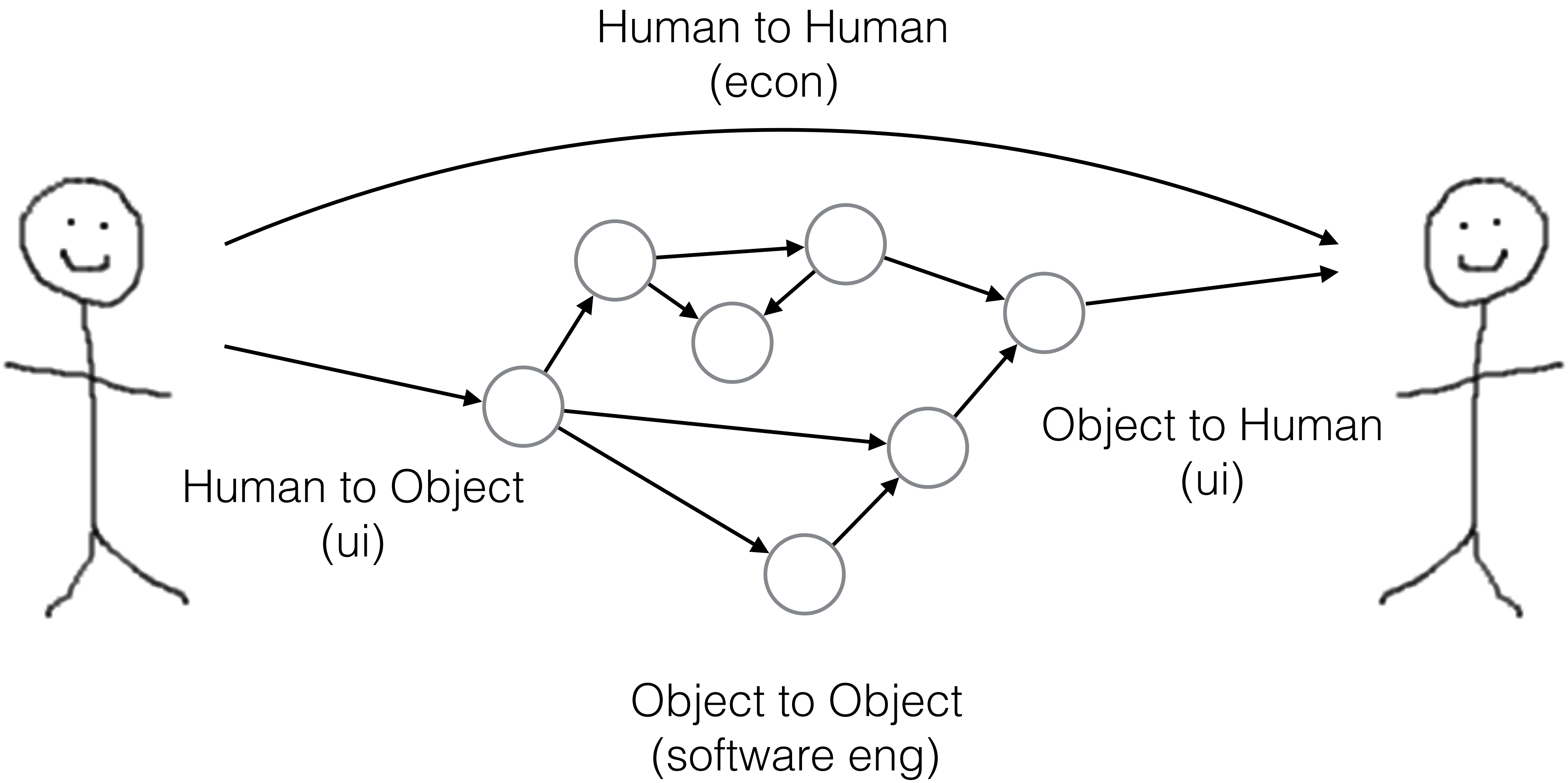
Principal-Agent relationships



“Give this to my Dad?”

Human to Human
(econ)

Principal-Agent relationships



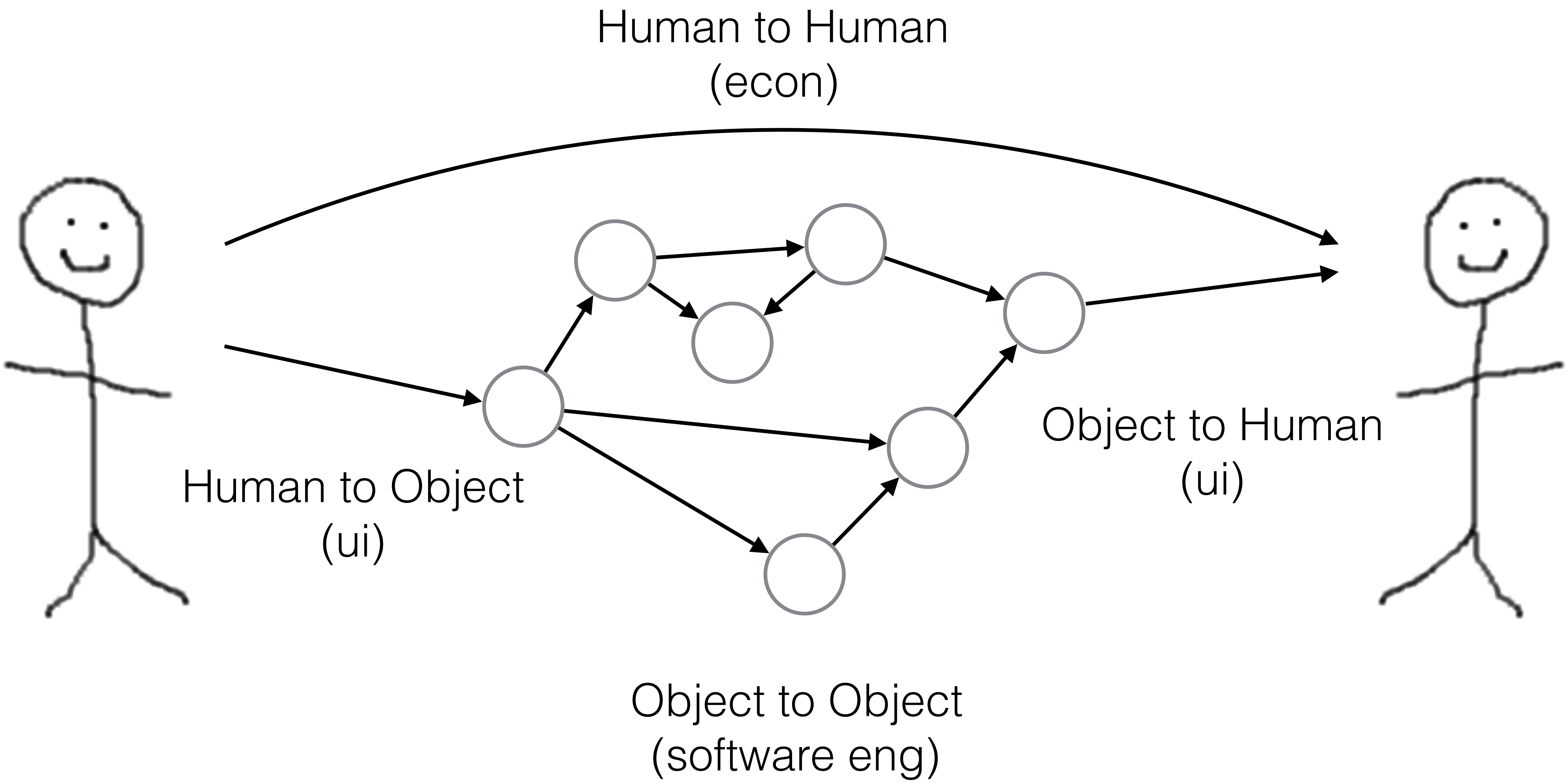
Human to Human
(econ)

Human to Object
(ui)

Object to Human
(ui)

Object to Object
(software eng)

Principal-Agent networks



Human to Human
(econ)

Human to Object
(ui)

Object to Human
(ui)

Object to Object
(software eng)

Definitions

From Econ:
A **Principal** sends a request to an **Agent**.
An **Agent** reacts to a request from a **Principal**.

Definitions

From Econ:
A Principal sends a request to an Agent.
An Agent reacts to a request from a Principal.

Incentive Alignment is when a (*human*) Principal or Agent uses **incentives** to induce the other's **intentions** to align with their own.

Definitions

From Econ:
A Principal sends a request to an Agent.
An Agent reacts to a request from a Principal.

Incentive Alignment is when a (*human*) Principal
or Agent uses incentives to induce
the other’s intentions to **align** with their own.

align: Compose well without interference.

Definitions

From Econ:
A Principal sends a request to an Agent.
An Agent reacts to a request from a Principal.

Incentive Alignment is when a (*human*) Principal
or Agent uses incentives to induce
the other's intentions to align with their own.

align: Compose well without interference.

From us:
Decision Alignment is when a Principal
or Agent uses *various tools* to make it more likely for
the other's *decisions and actions* to align with their own.

Definitions

From us:
Decision Alignment is when a Principal or Agent uses *various tools* to make it more likely for the other's *decisions and actions* to align with their own.

Definitions

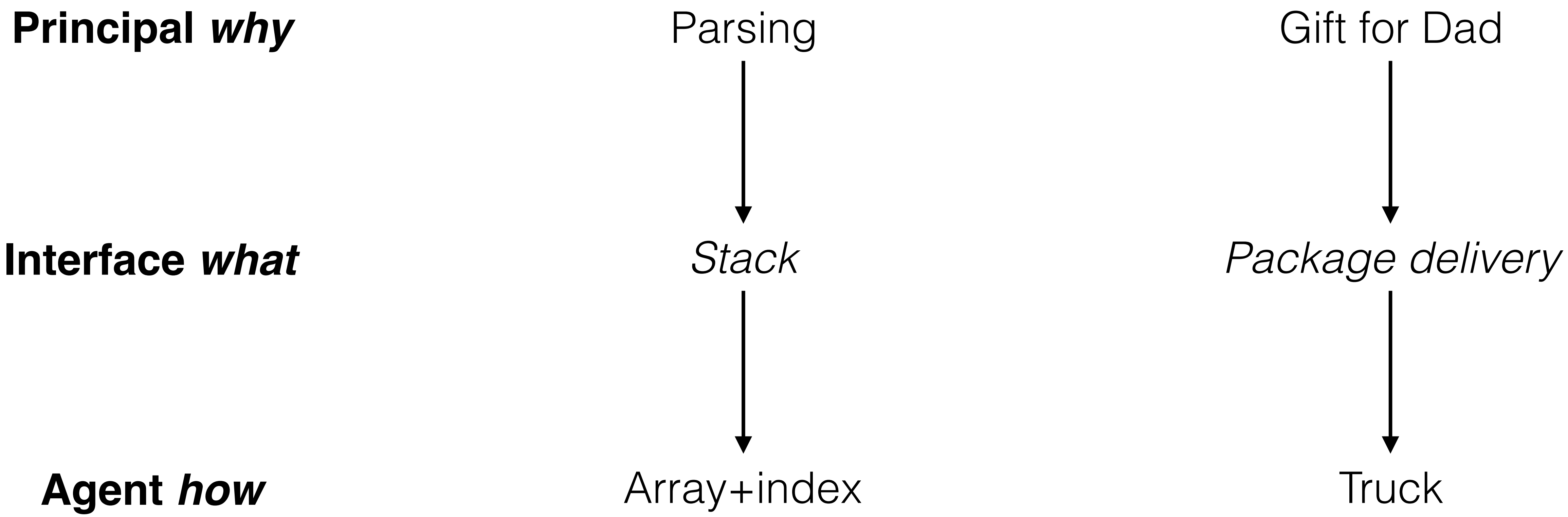
From us:
Decision Alignment is when a *Principal or Agent* uses various tools to make it more likely for *the other’s* decisions and actions to align with their own.

This talk:
Decision Alignment is when a *Principal* uses various tools to make it more likely for *the Agent’s* decisions and actions to align with its own.

Information Hiding Benefits

Compose specialized knowledge

Only Principal knows ***why***
Only Agent knows ***how***
Shared knowledge burden: only what's needed for request
Minimize cascading changes



Principal *why* Parsing Gift for Dad

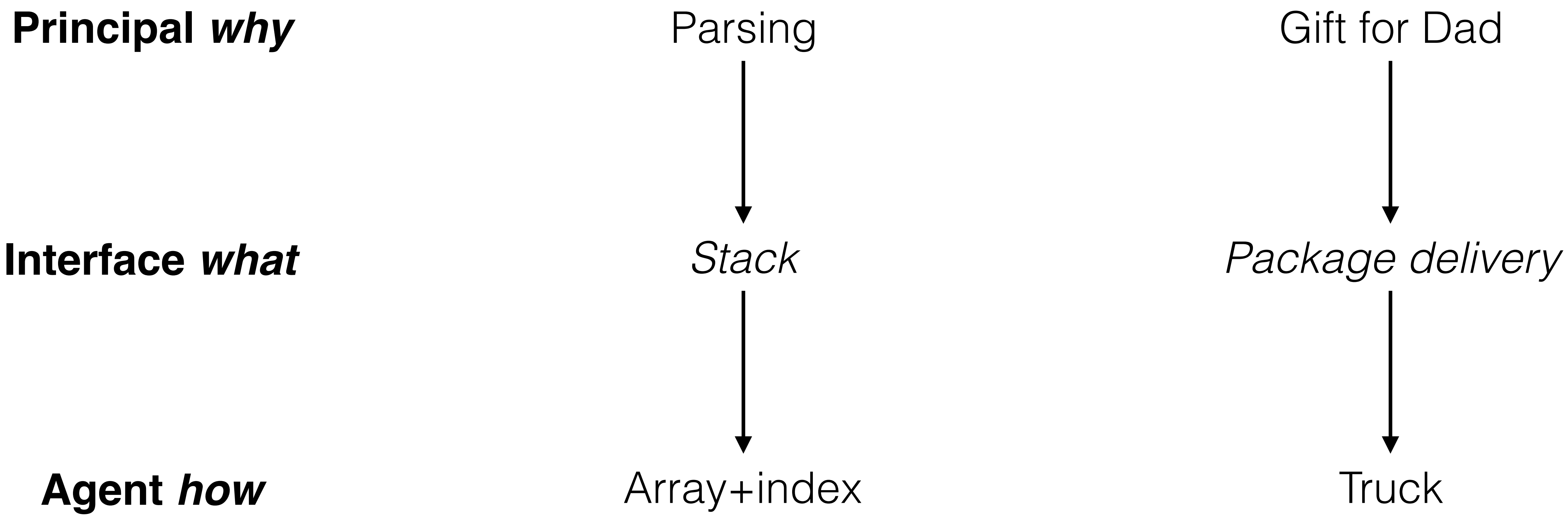
Interface *what* *Stack* *Package delivery*

Agent *how* Array+index Truck

Hidden Information Hazards

The principal-agent problem

Pre: Can the agent do what I want?
Request: Will the agent try to do what I want?
Post: Is the agent doing what I want?

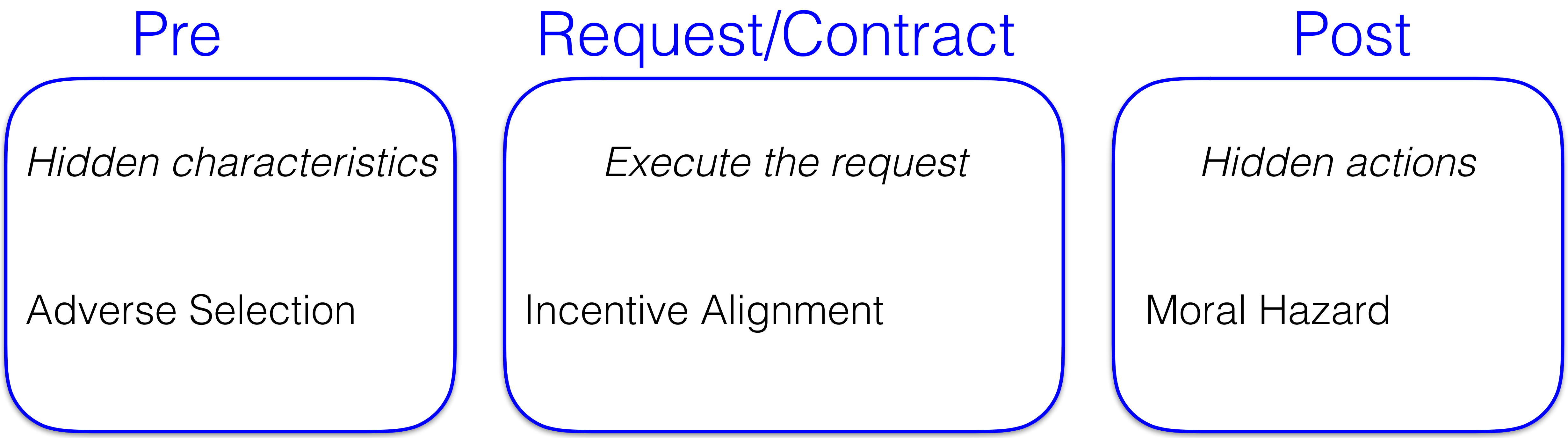


Principal *why* Parsing Gift for Dad

Interface *what* *Stack* *Package delivery*

Agent *how* Array+index Truck

The Principal-Agent Loop



Pre

Hidden characteristics

Adverse Selection

Request/Contract

Execute the request

Incentive Alignment

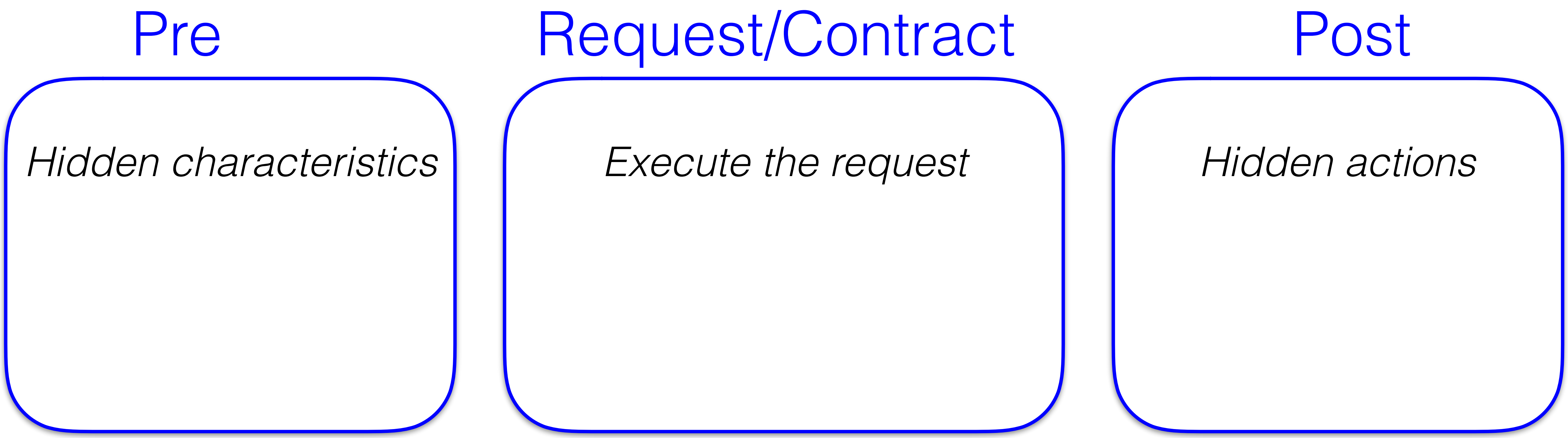
Post

Hidden actions

Moral Hazard

Econon: intentional misbehavior

The Principal-Agent Loop



Pre

Hidden characteristics

Request/Contract

Execute the request

Post

Hidden actions

Econon: intentional misbehavior

CS: accidental misbehavior

The Principal-Agent Loop

Pre

Hidden characteristics

Select agent.
Screening.
Agent signals.

Inspect internals.
Abilities, limits.

Request/Contract

Execute the request

Allow actions.
Scope of authority.

Explain request.
What the agent is supposed to do.

Reward cooperation.
If agent does that.

Post

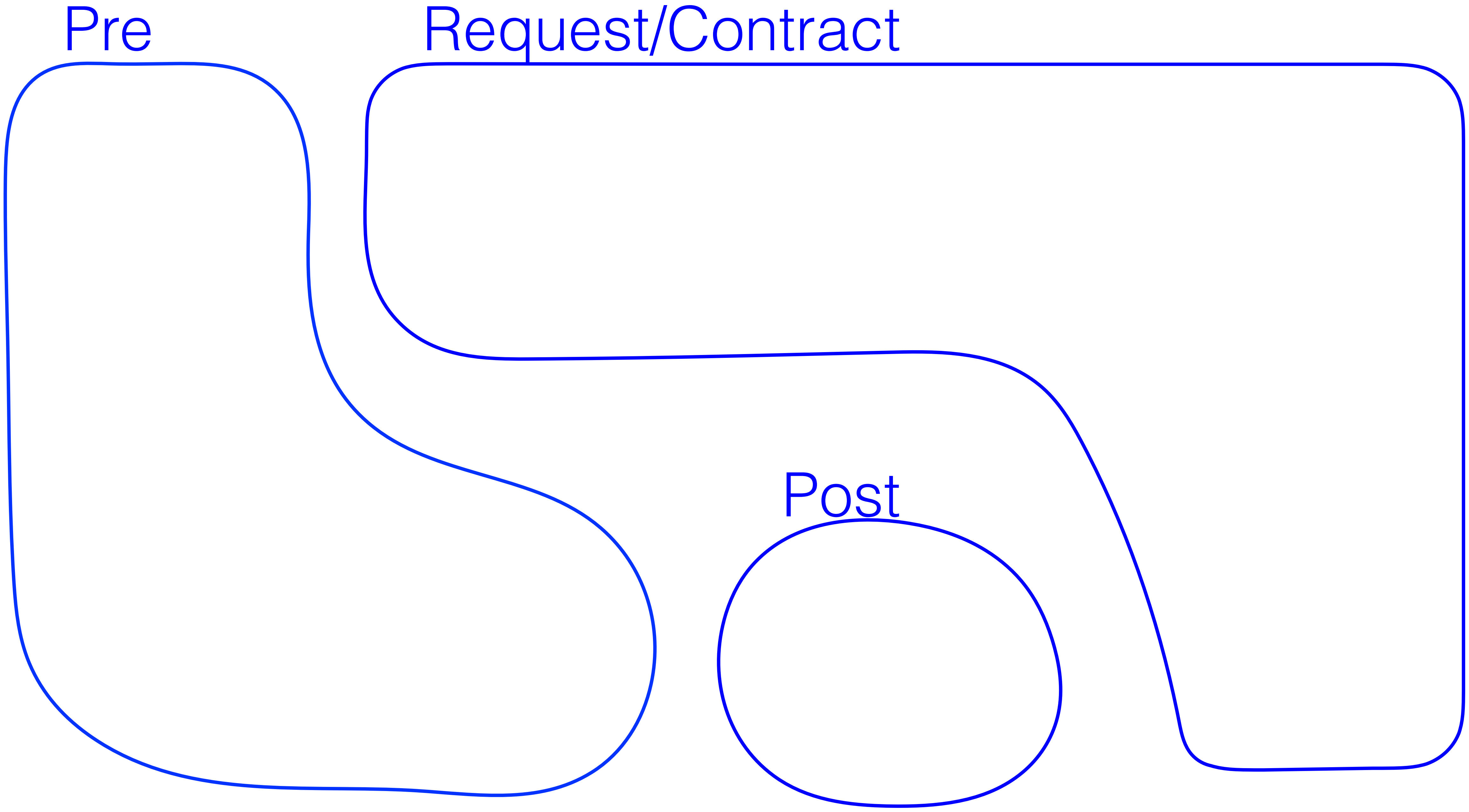
Hidden actions

Monitor effects.
What agent is doing, or did.

Feedback to guide future selection.

The Principal-Agent Loop

Only loosely ordered



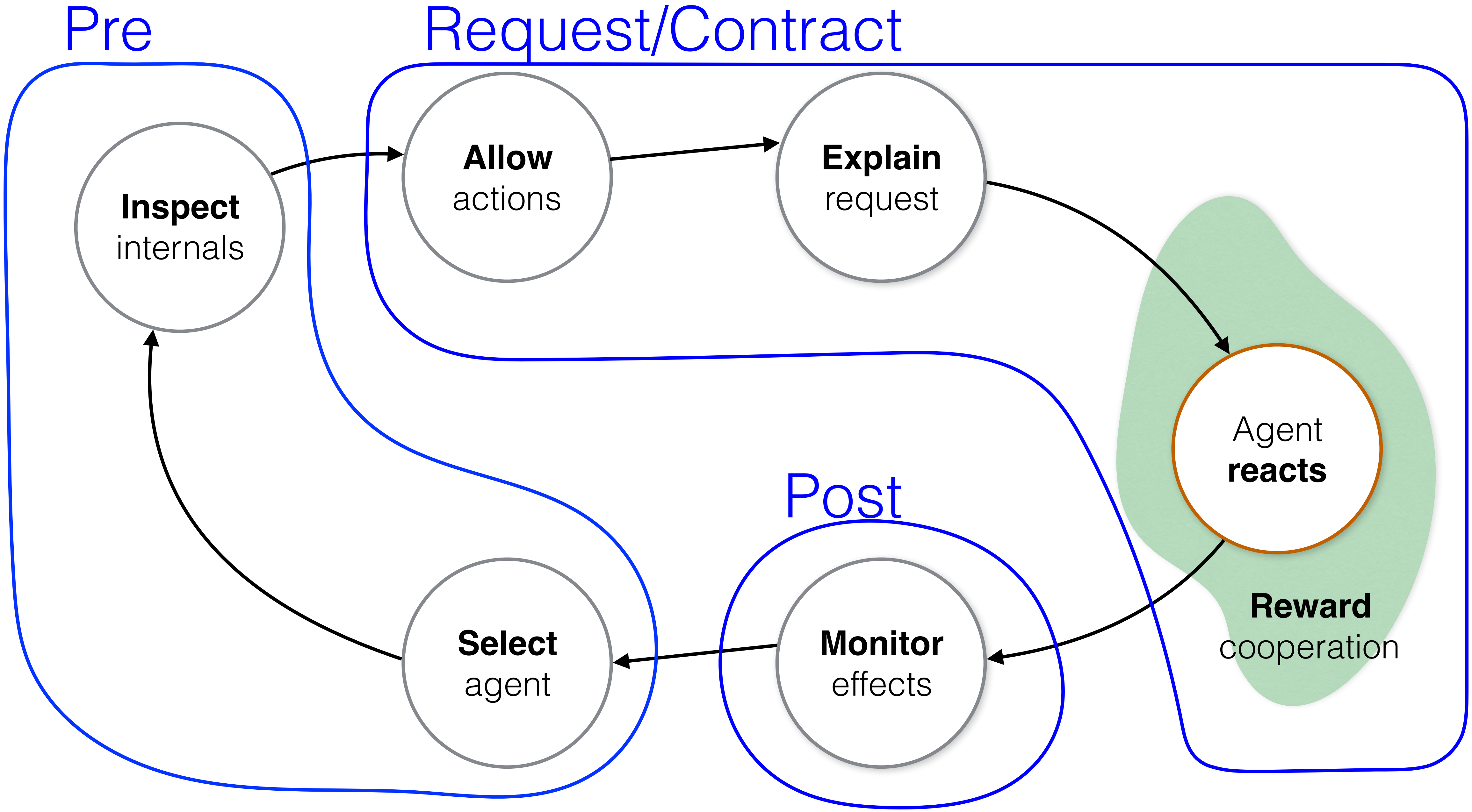
Pre

Request/Contract

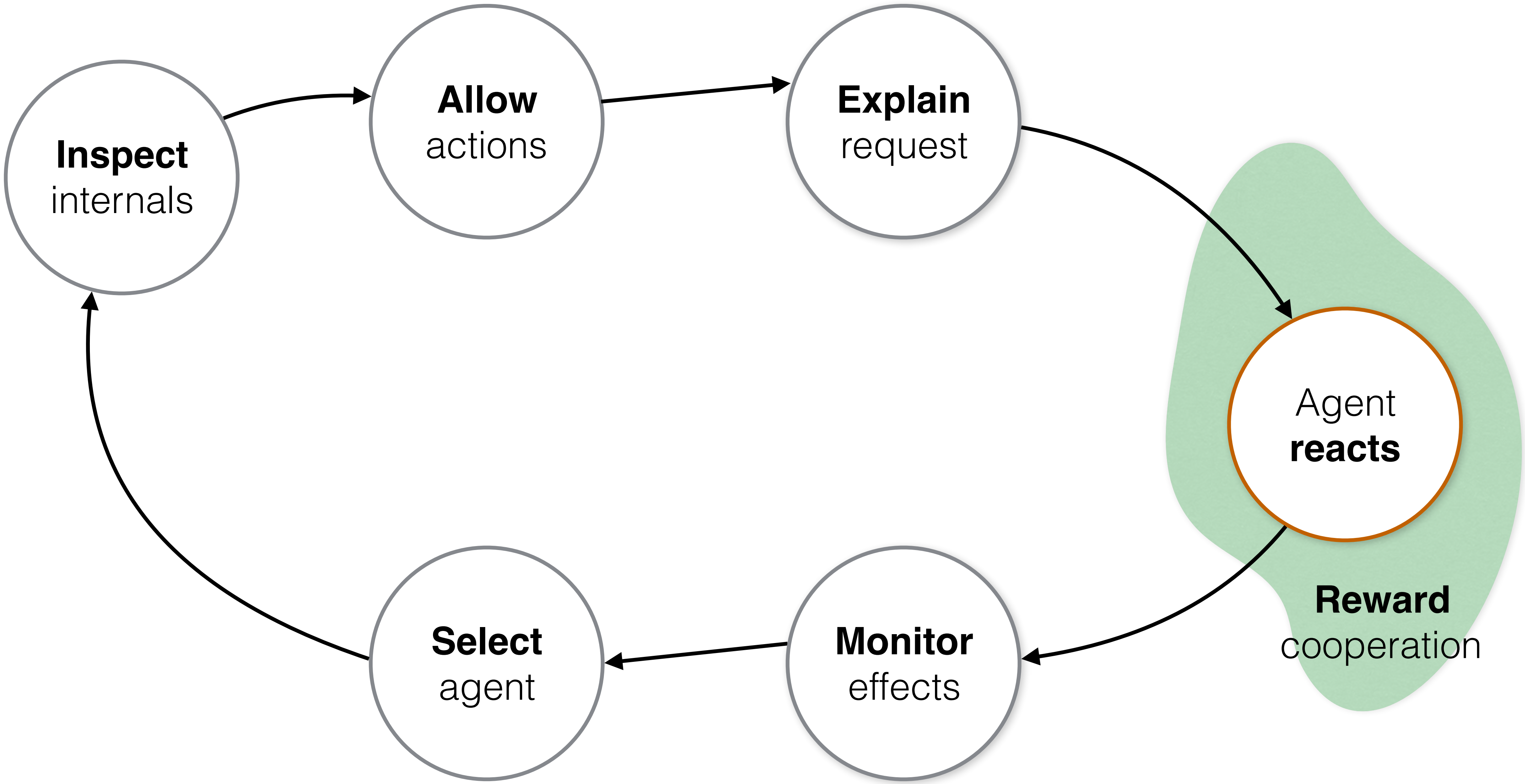
Post

The Principal-Agent Loop

Only loosely ordered



The Principal-Agent Loop



Inspect internals

Allow actions

Explain request

Agent **reacts**

Reward cooperation

Monitor effects

Select agent

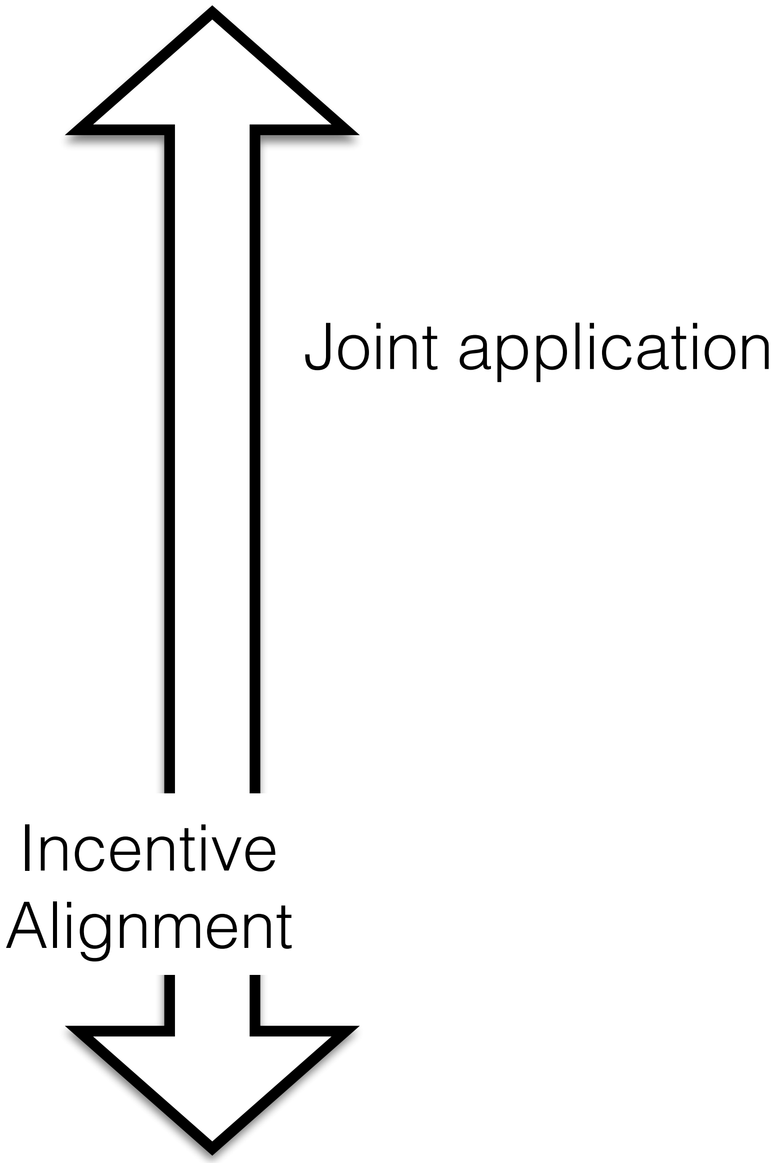
From Incentive Alignment …

	Human to Human	Human to Object	Object to Object	Object to Human
Select agent				
Inspect internals				
Allow actions				
Explain request				
Reward cooperation	Incentive Alignment			
Monitor effects				

… to Principal-Agent …

Recognize synergies

	Human to Human	Human to Object	Object to Object	Object to Human
Select agent				
Inspect internals				
Allow actions				
Explain request				
Reward cooperation				
Monitor effects				



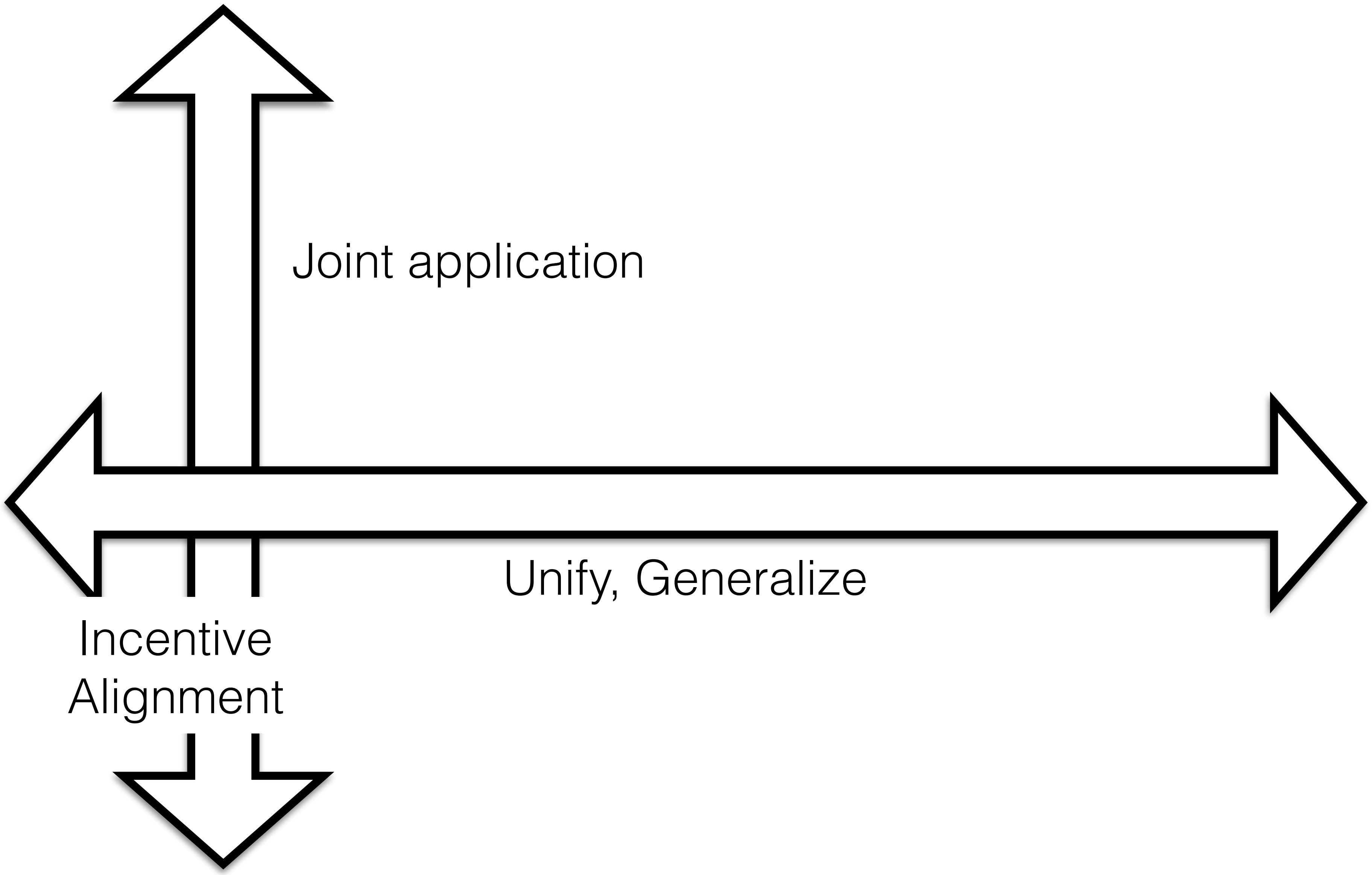
Joint application

Incentive Alignment

… to Decision Alignment

Recognize commonalities

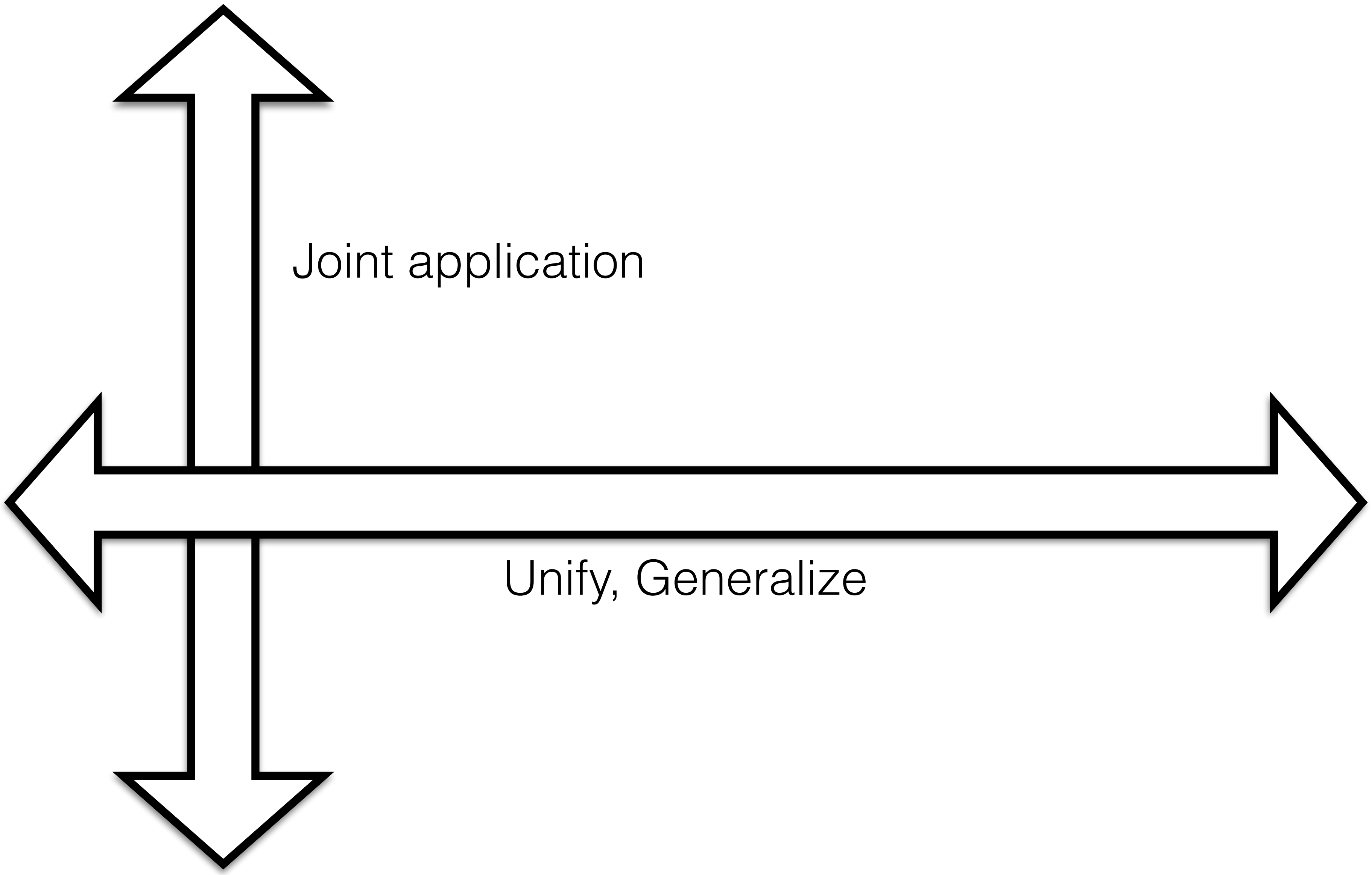
	Human to Human	**Human to Object**	**Object to Object**	**Object to Human**
Select agent				
Inspect internals				
Allow actions				
Explain request				
Reward cooperation				
Monitor effects				



… to Decision Alignment

Recognize commonalities

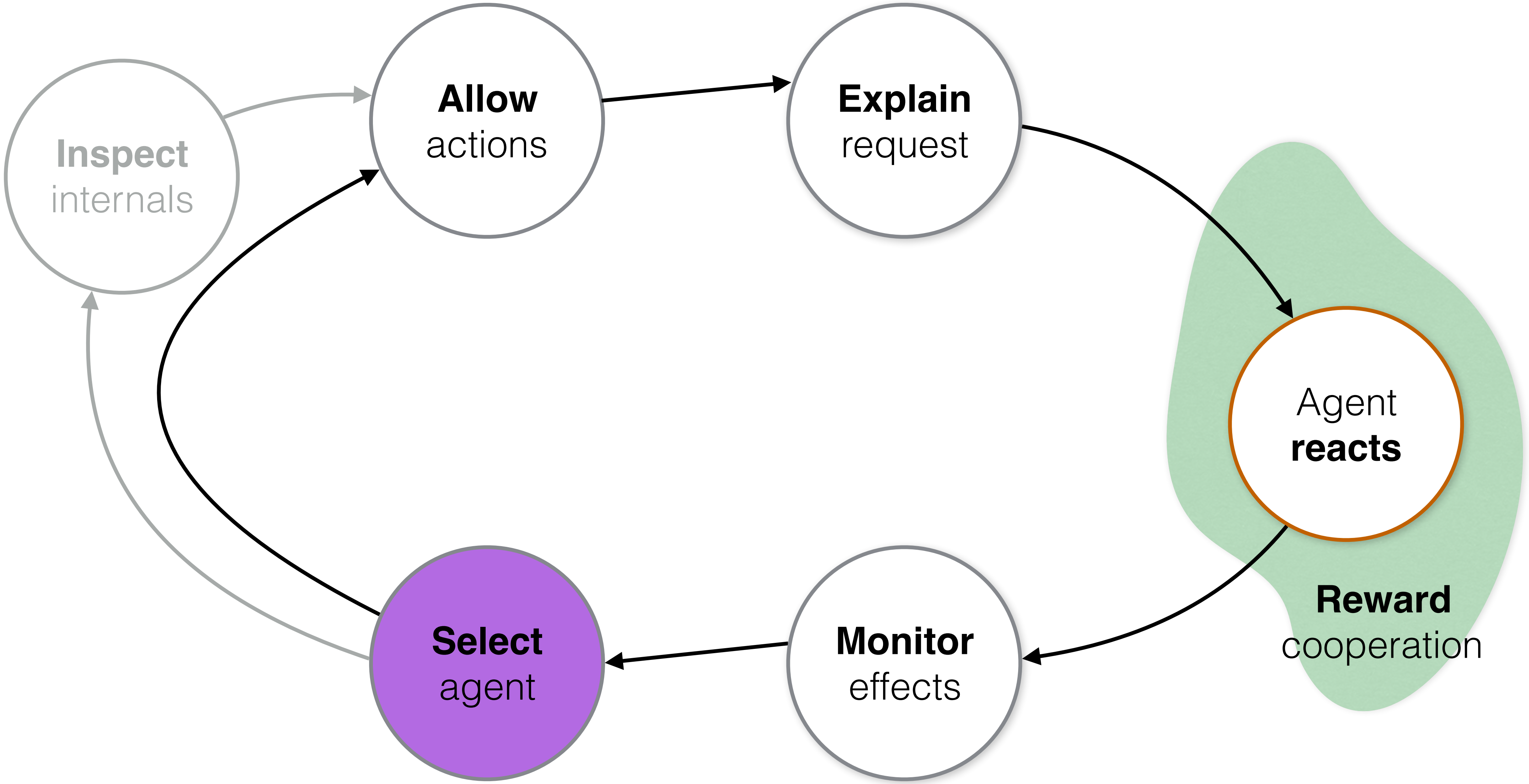
	Human to Human	**Human to Object**	**Object to Object**	**Object to Human**
Select agent				
Inspect internals				
Allow actions				
Explain request				
Reward cooperation				
Monitor effects				



ups
Sincronizziamo il mondo del commercio
800 877 877
www.ups.com
GV.30697
COVANCE

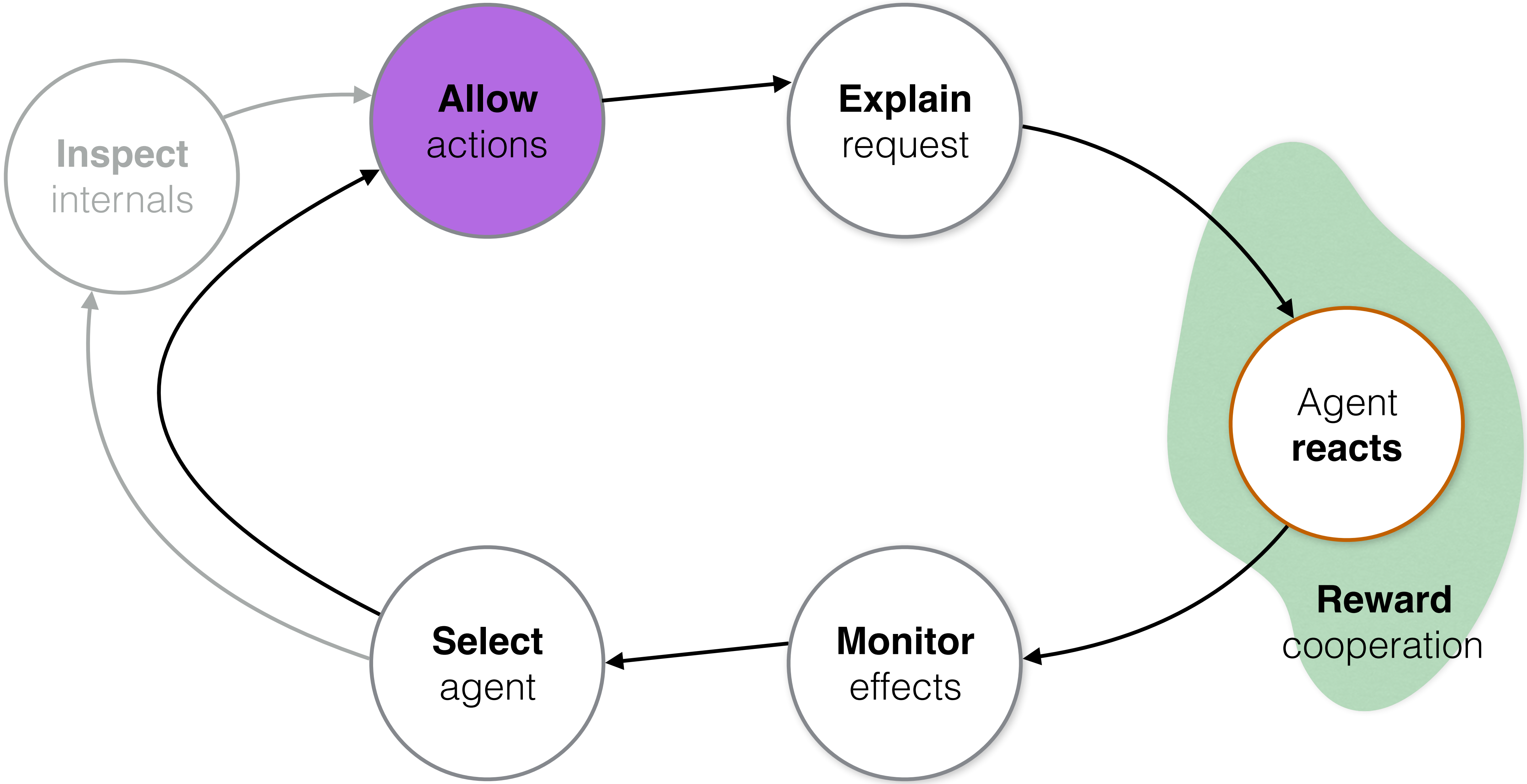
Package Delivery

Fit, Reputation



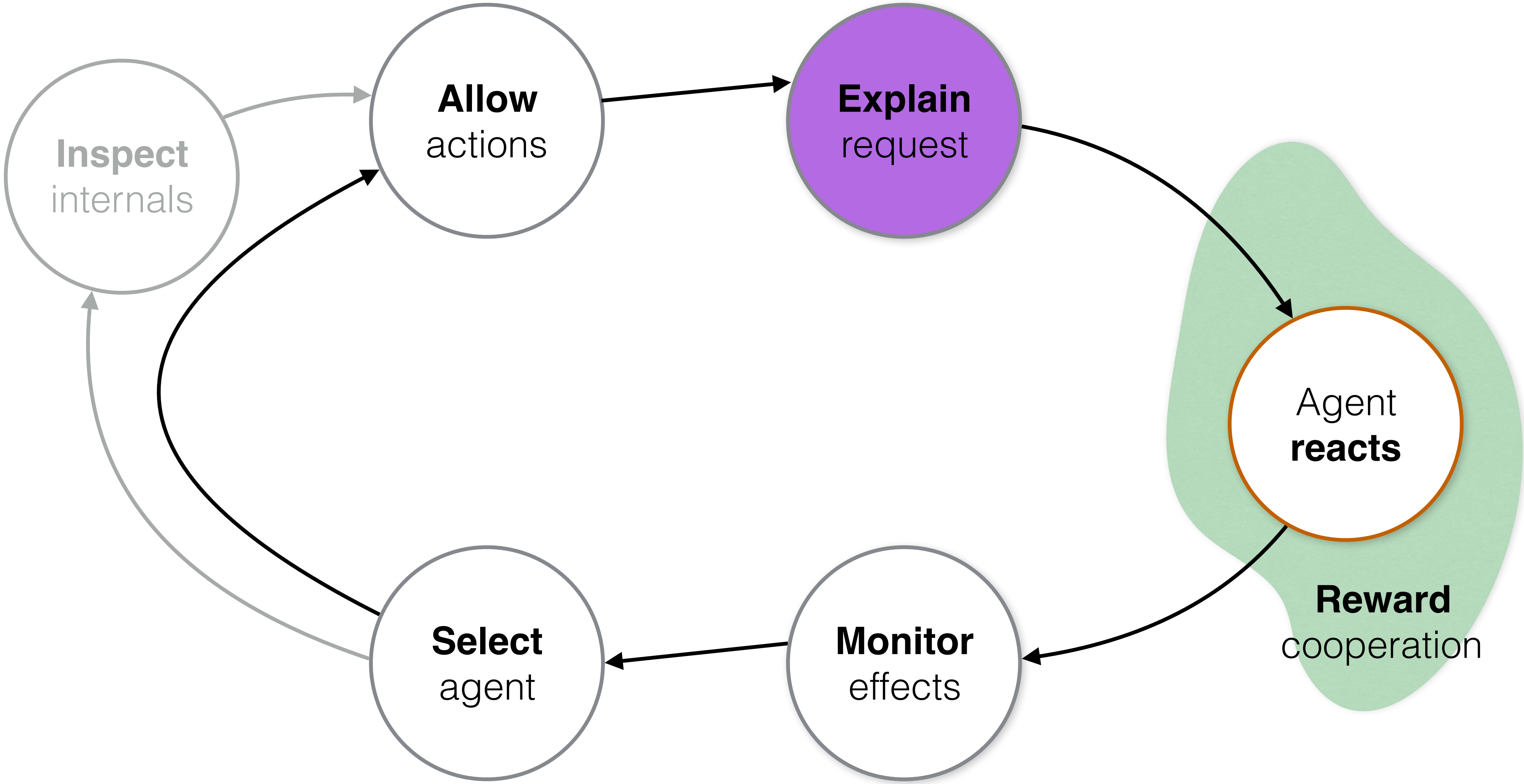
Package Delivery

Hand over the package



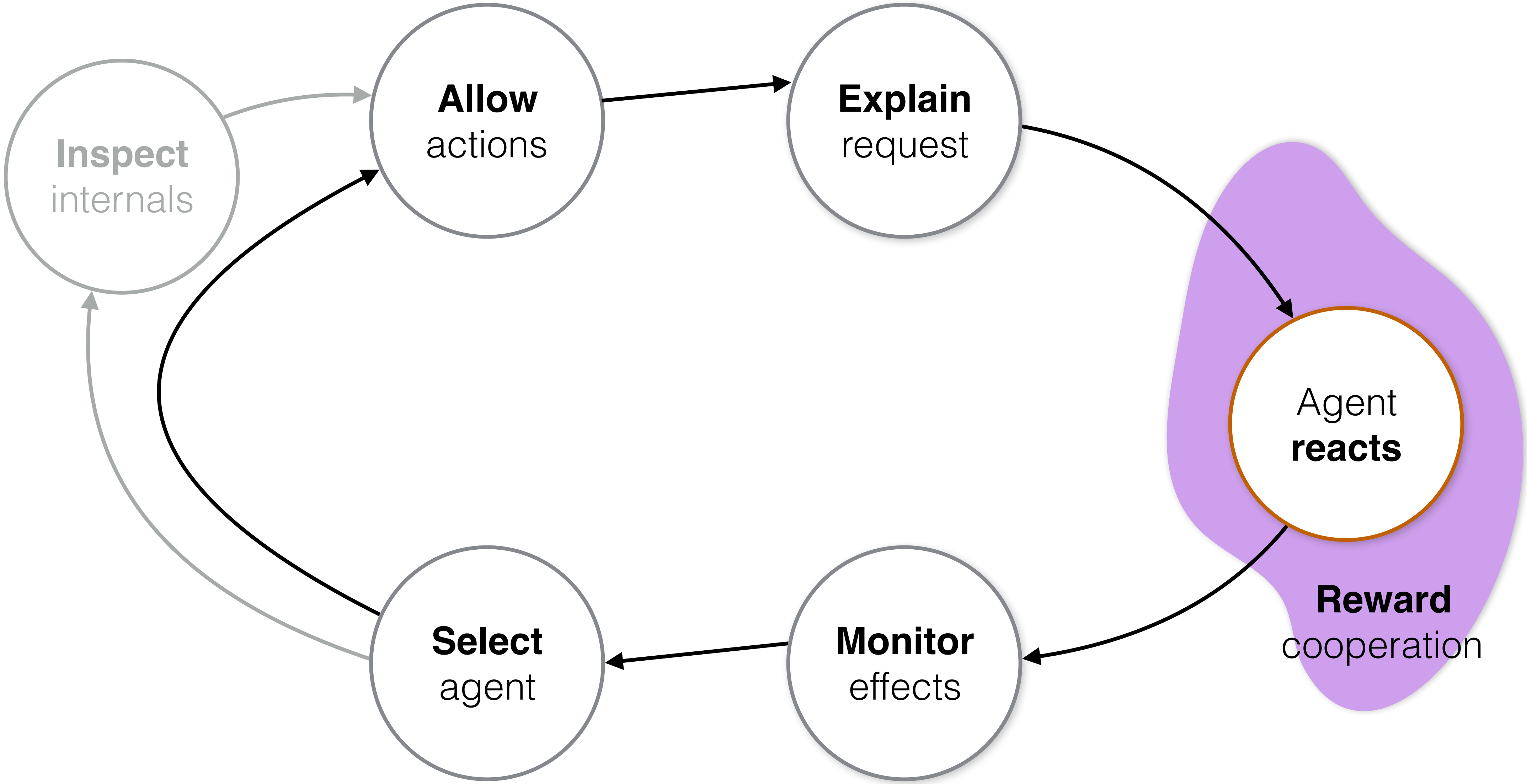
Package Delivery

Delivery address, instructions



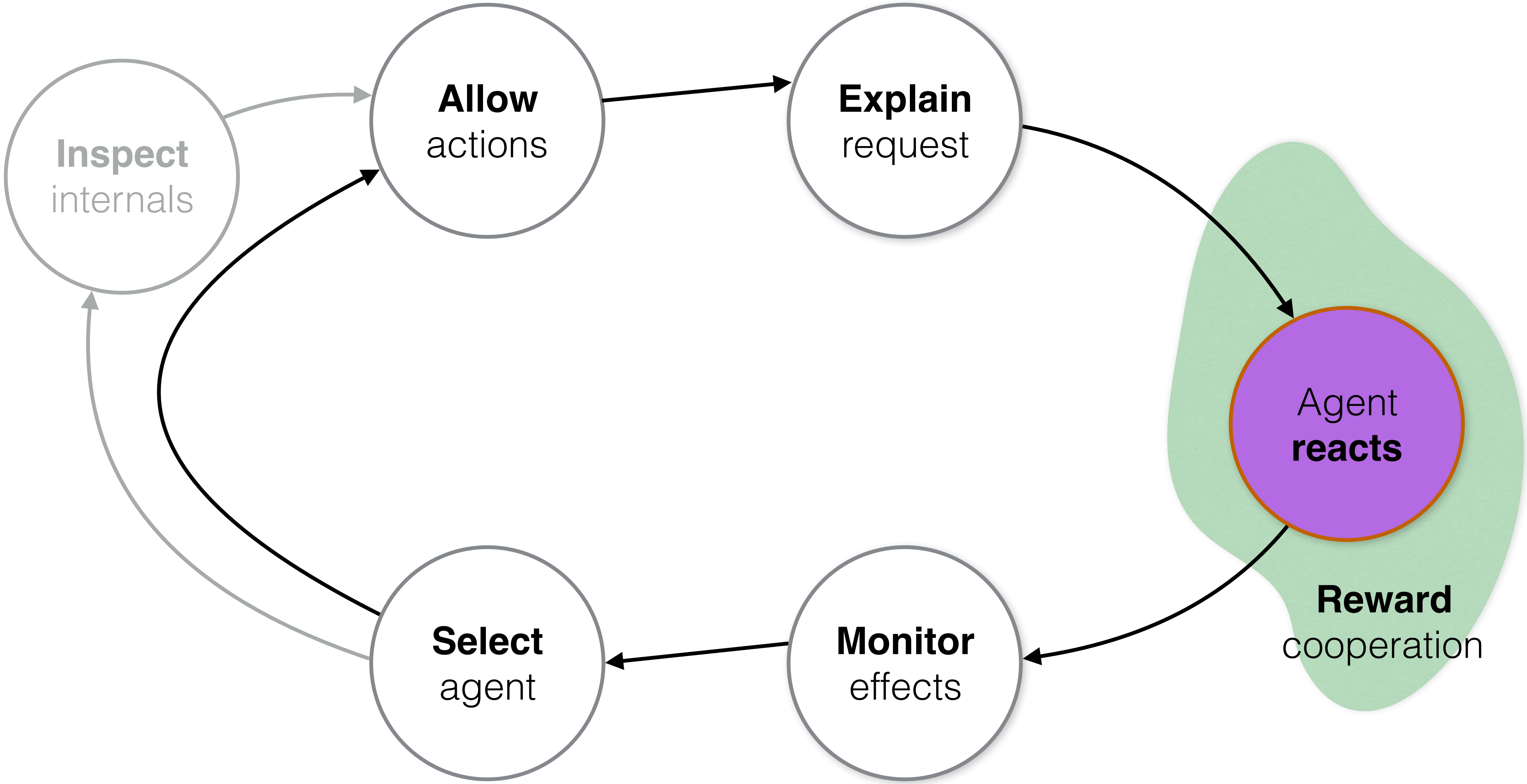
Package Delivery

Pay



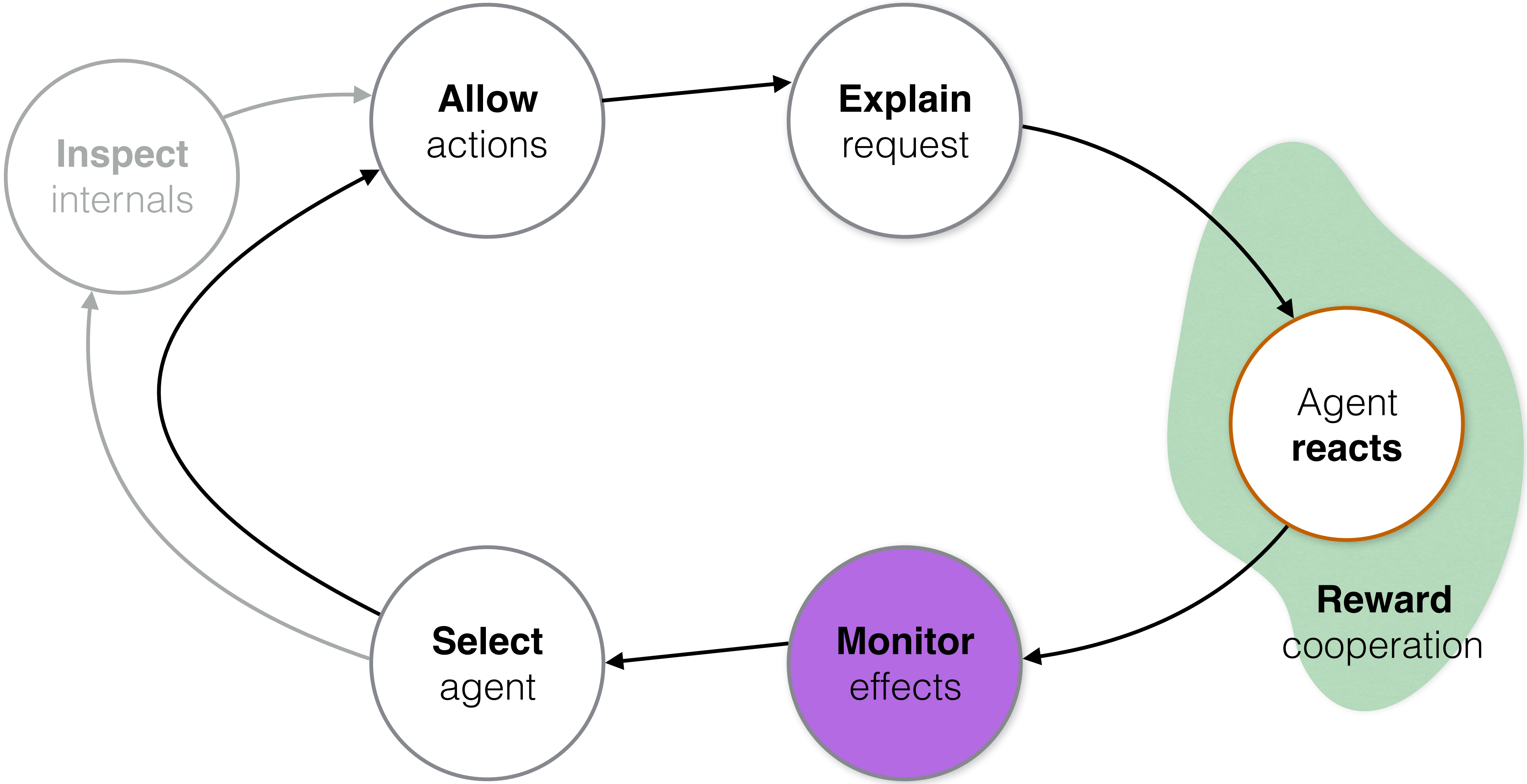
Package Delivery

Hope and pray?



Package Delivery

Track, Return receipt



Inspect internals

Allow actions

Explain request

Agent **reacts**

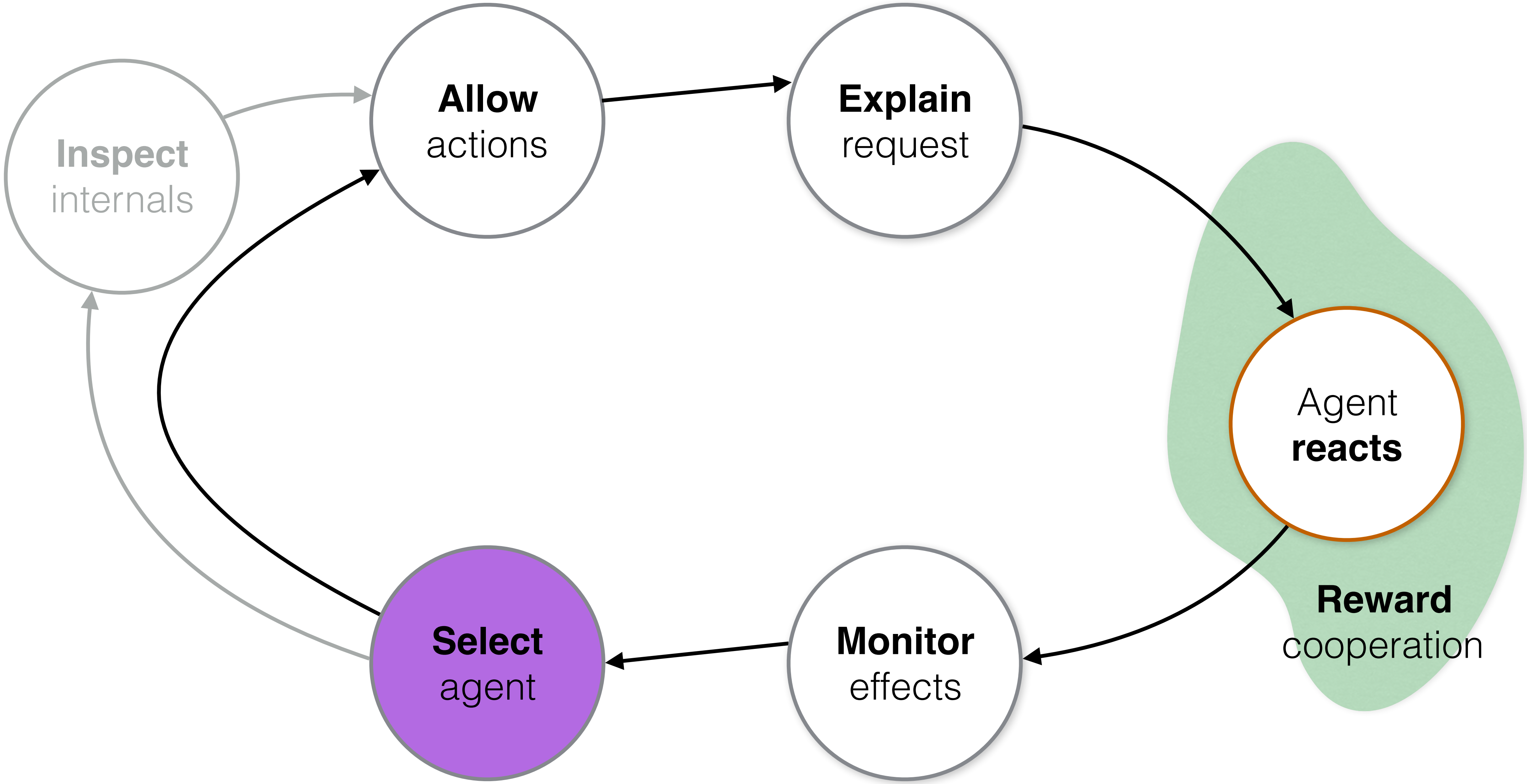
Reward cooperation

Monitor effects

Select agent

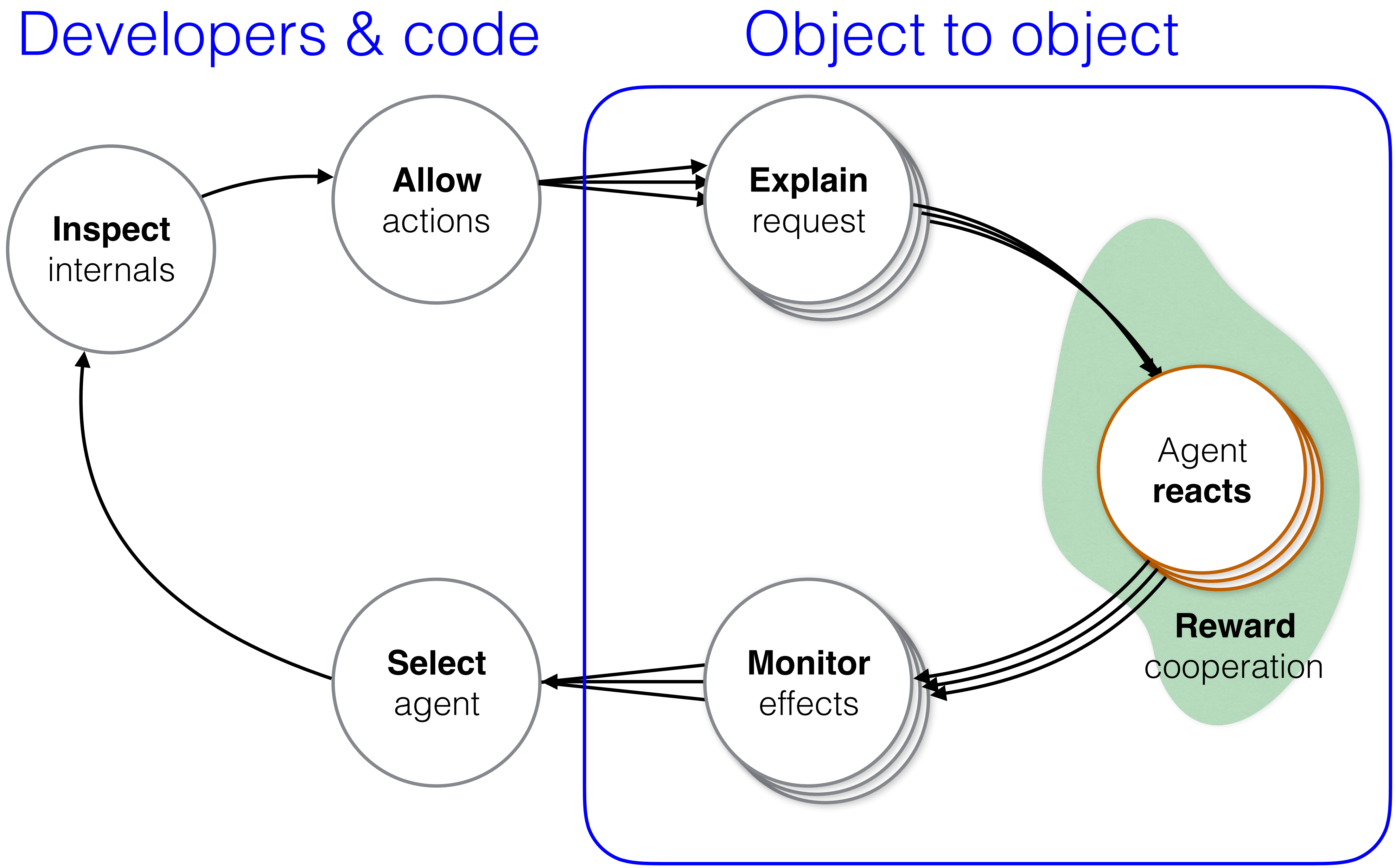
Package Delivery

Reputation feedback, Rating



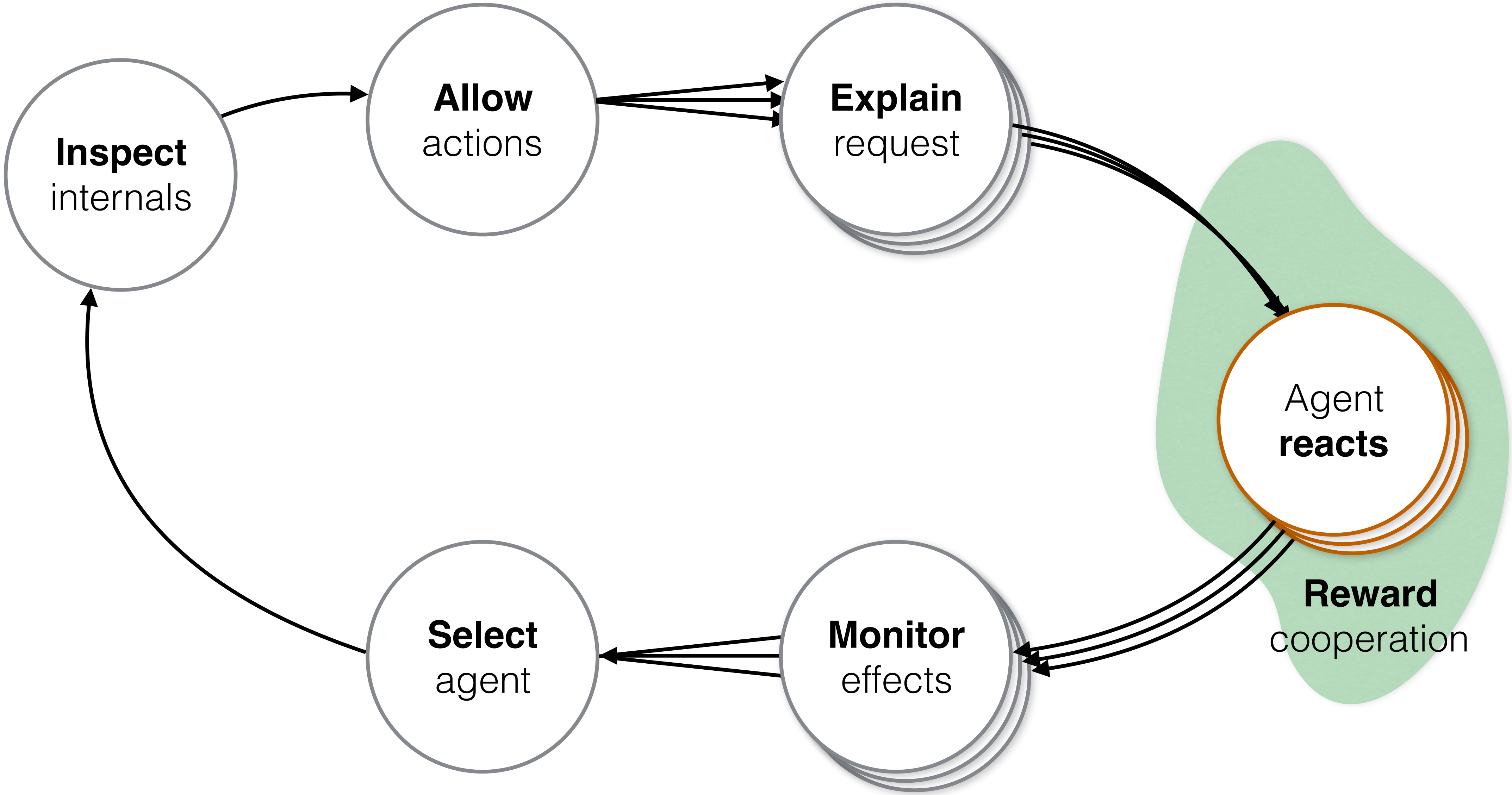
Internal Software Development

Static *vs.* Dynamic



Internal Software Development

De-emphasize Rewards



Inspect internals

Allow actions

Explain request

Agent **reacts**

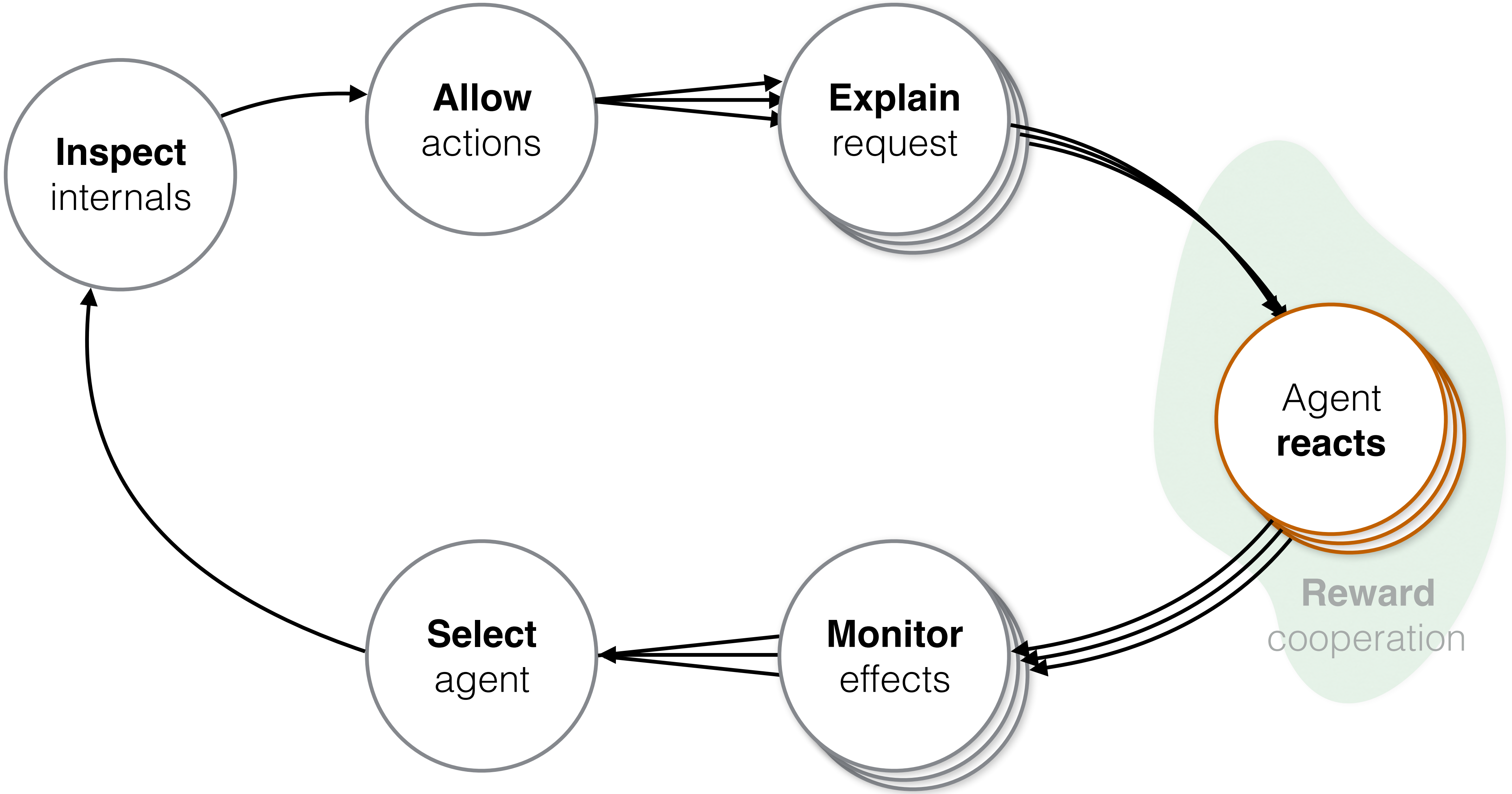
Reward cooperation

Monitor effects

Select agent

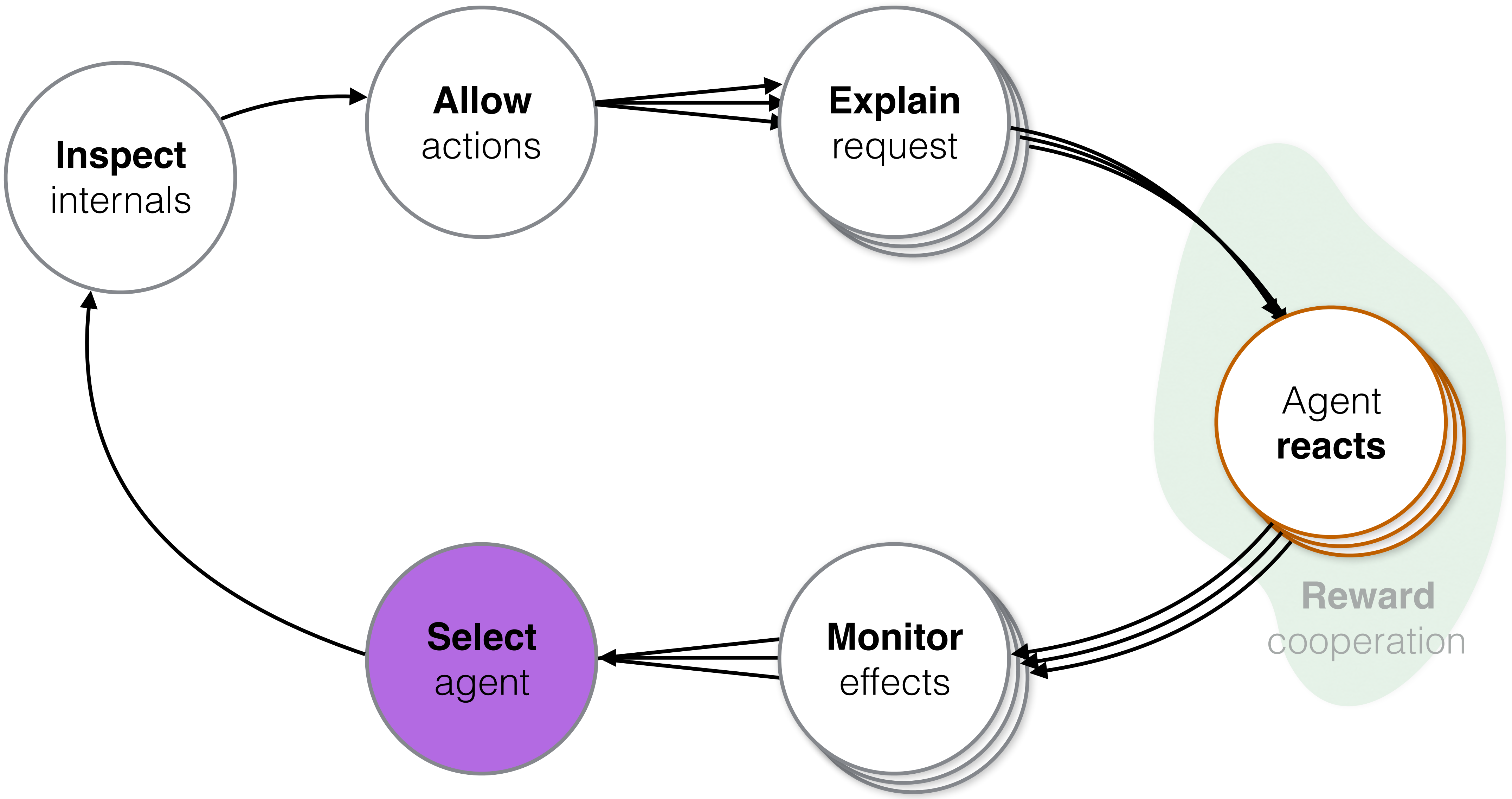
Internal Software Development

De-emphasize Rewards



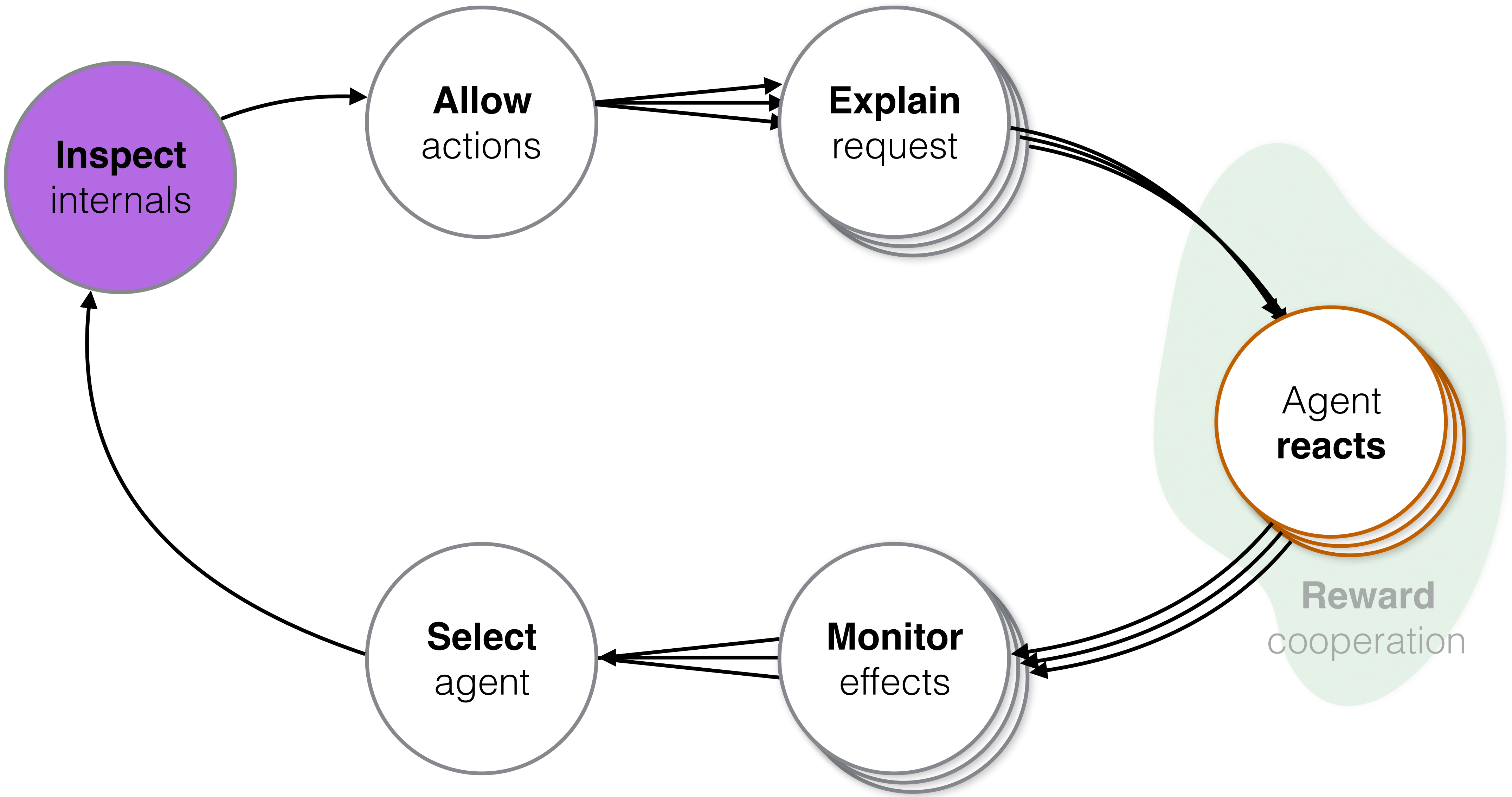
Internal Software Development

Hire the best. Find libraries.



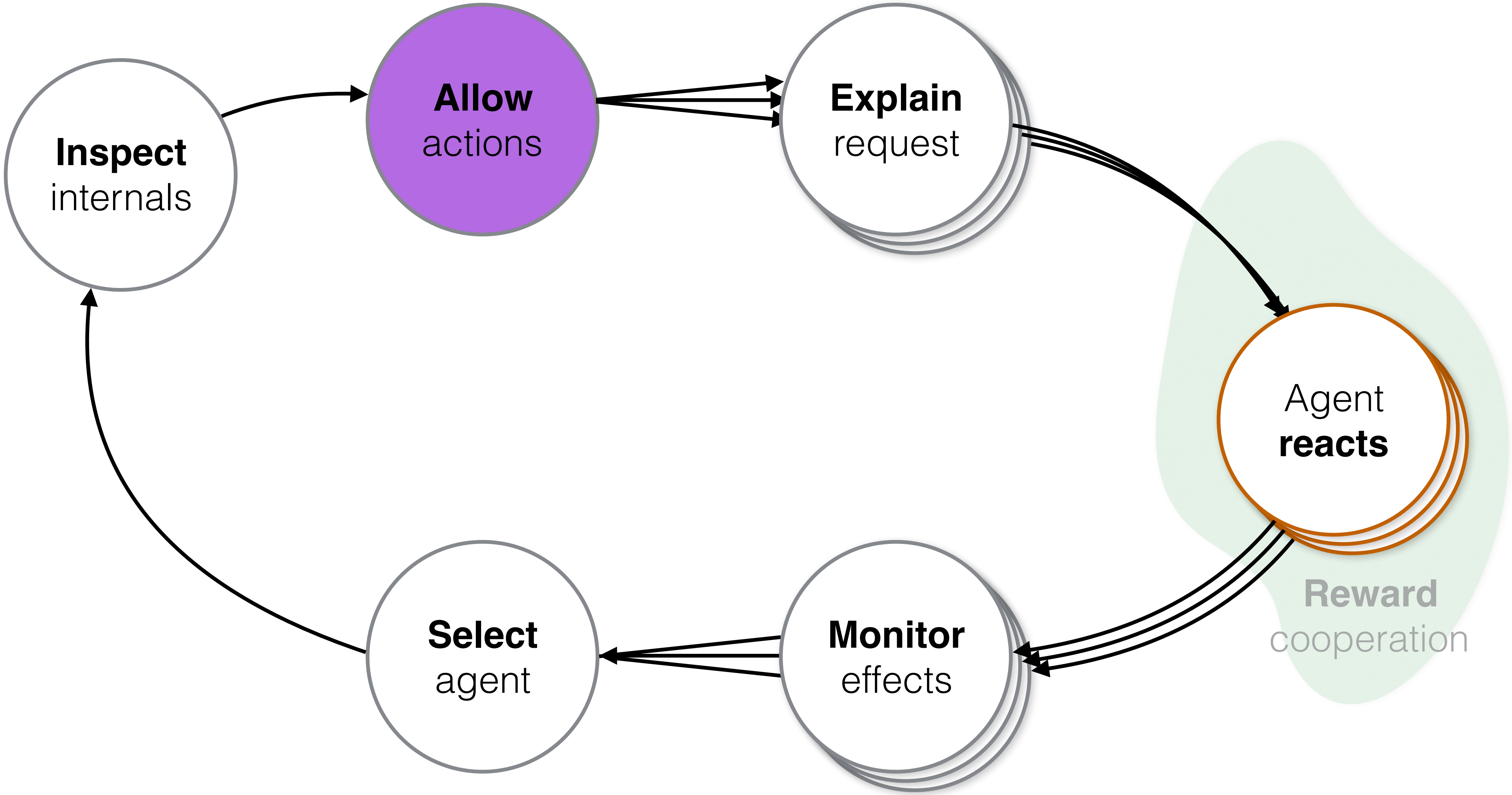
Internal Software Development

Code reviews. Some static checking.



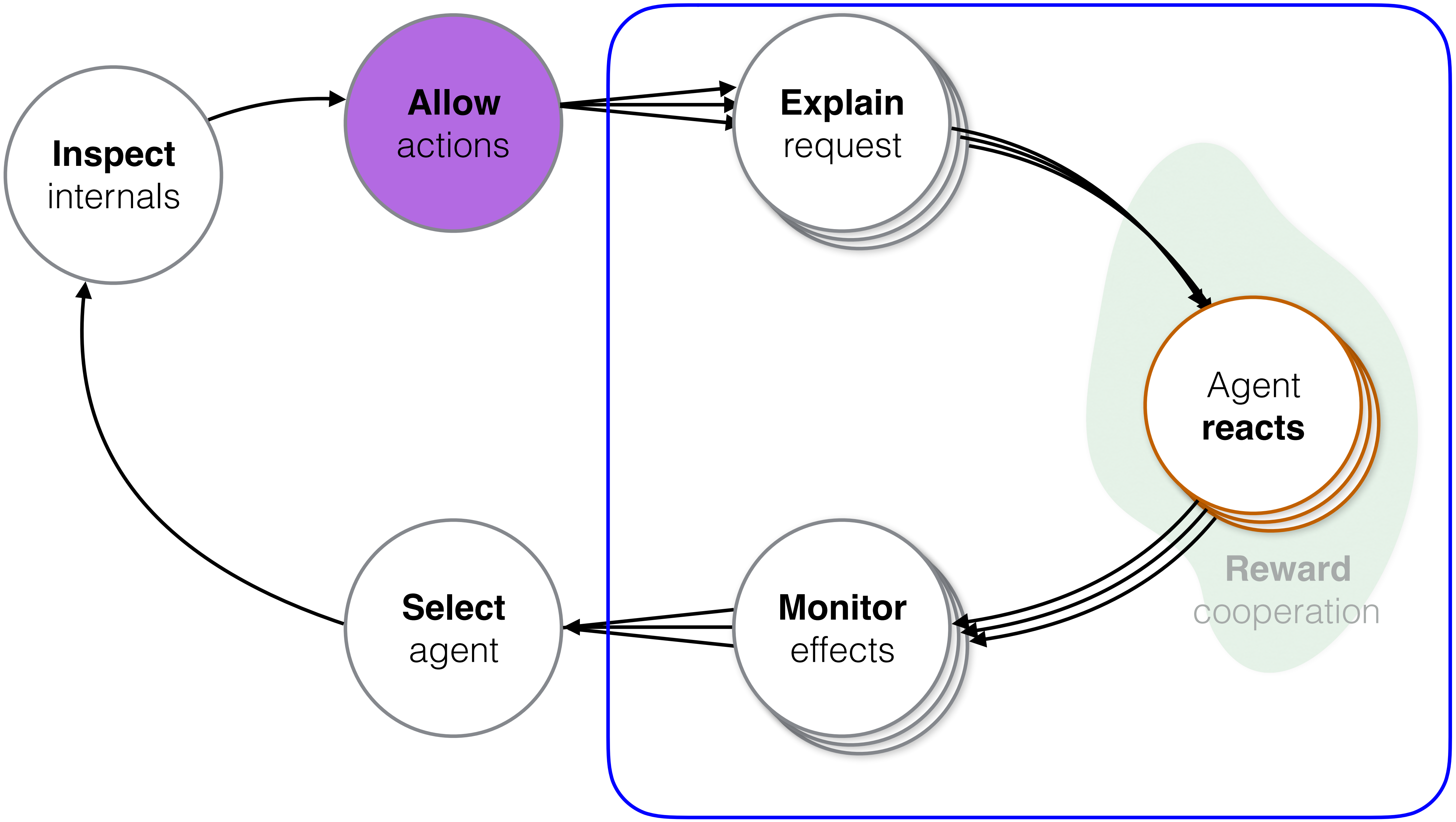
Internal Software Development

All user's authority.



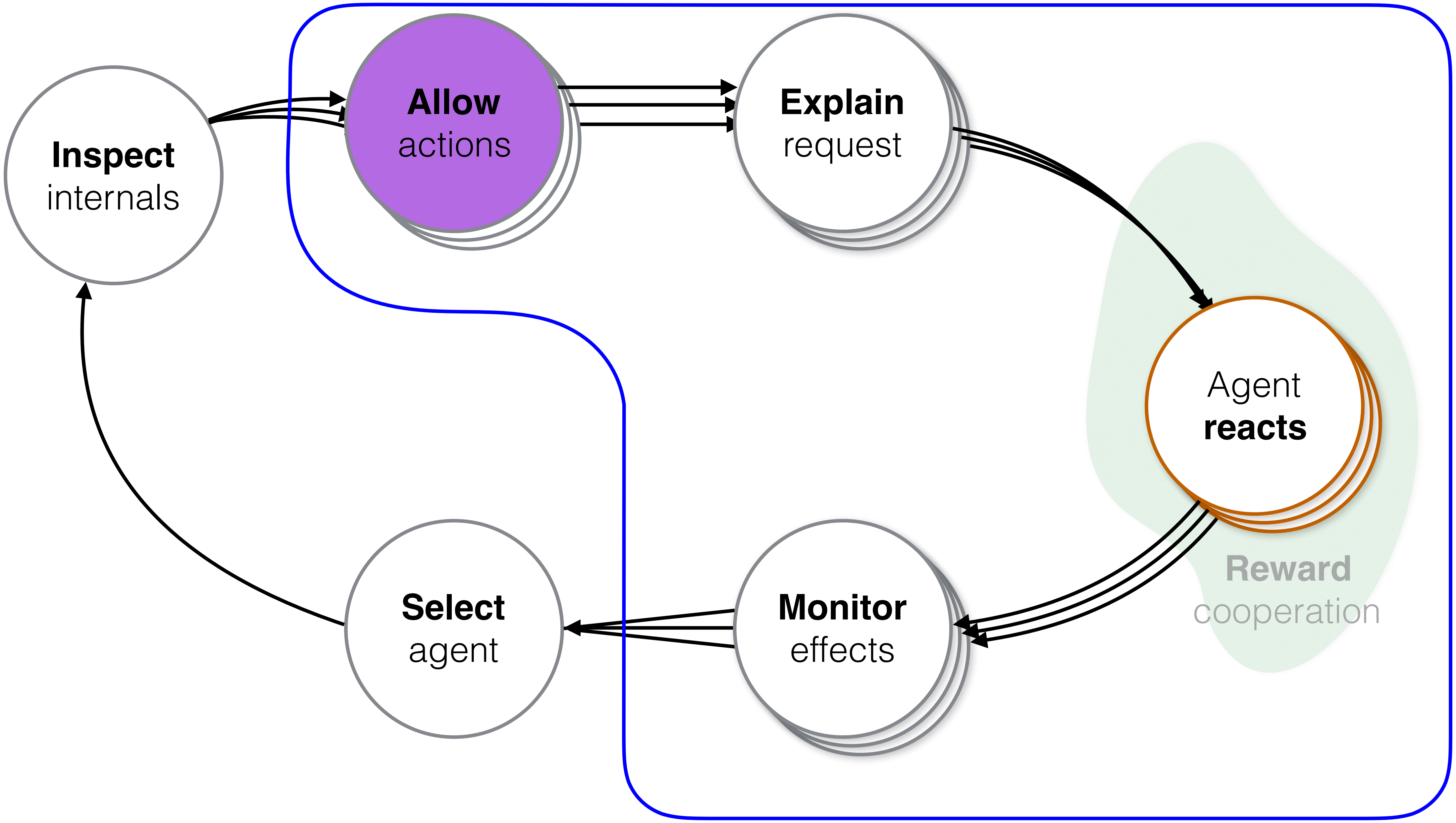
Internal Software Development

All user's authority.



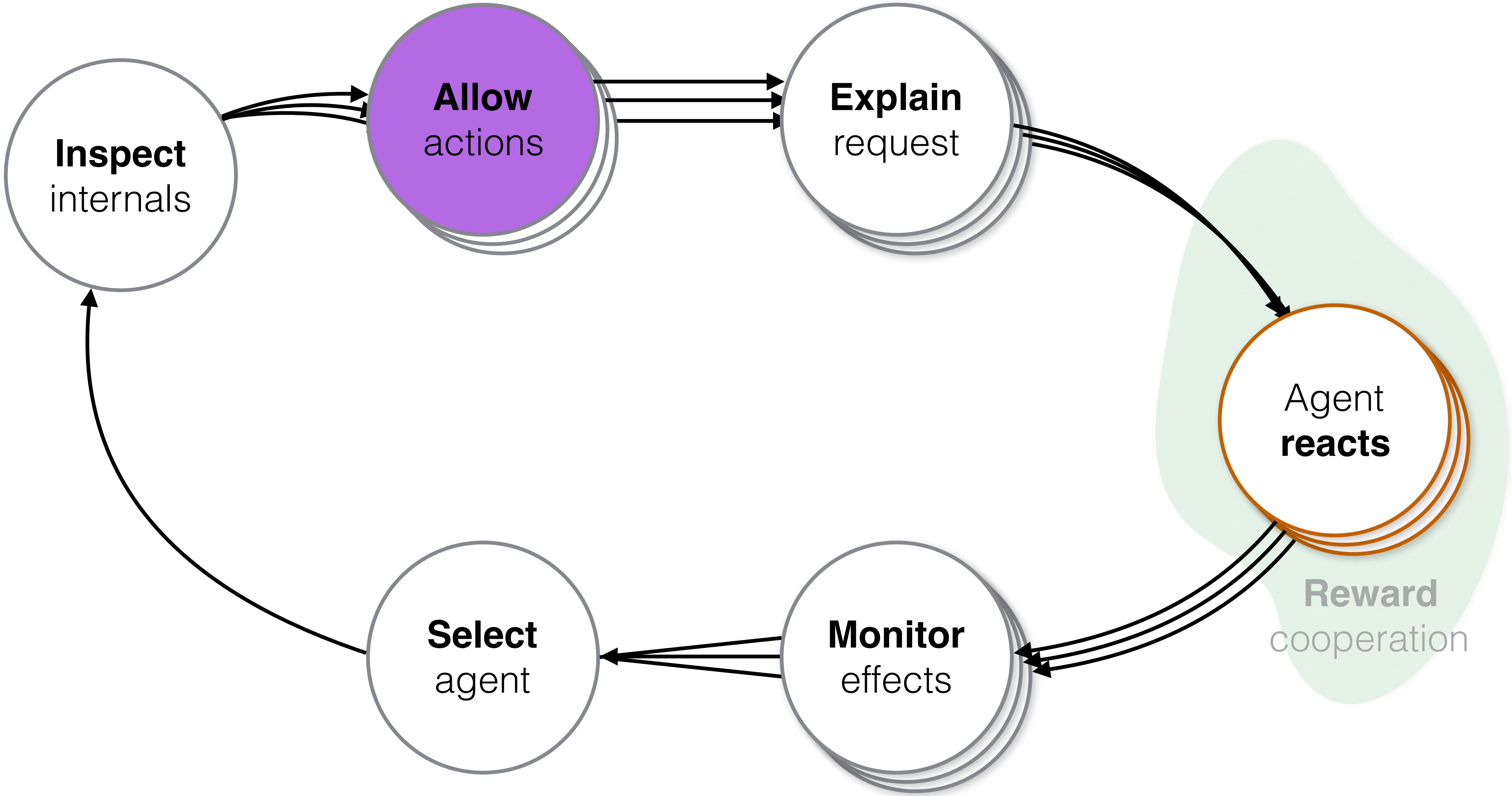
Internal Software Development

Rights per request



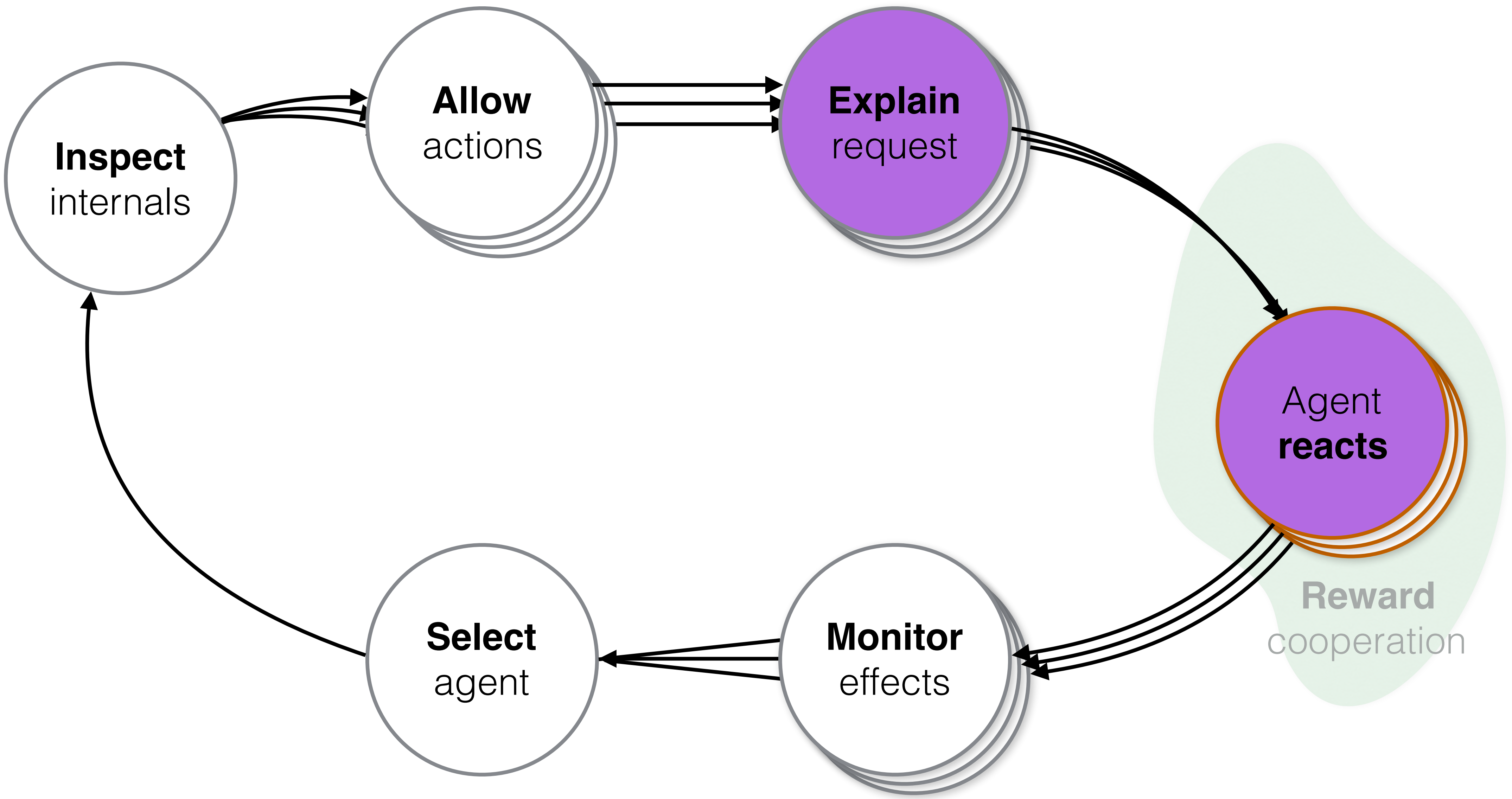
Internal Software Development

Rights per request



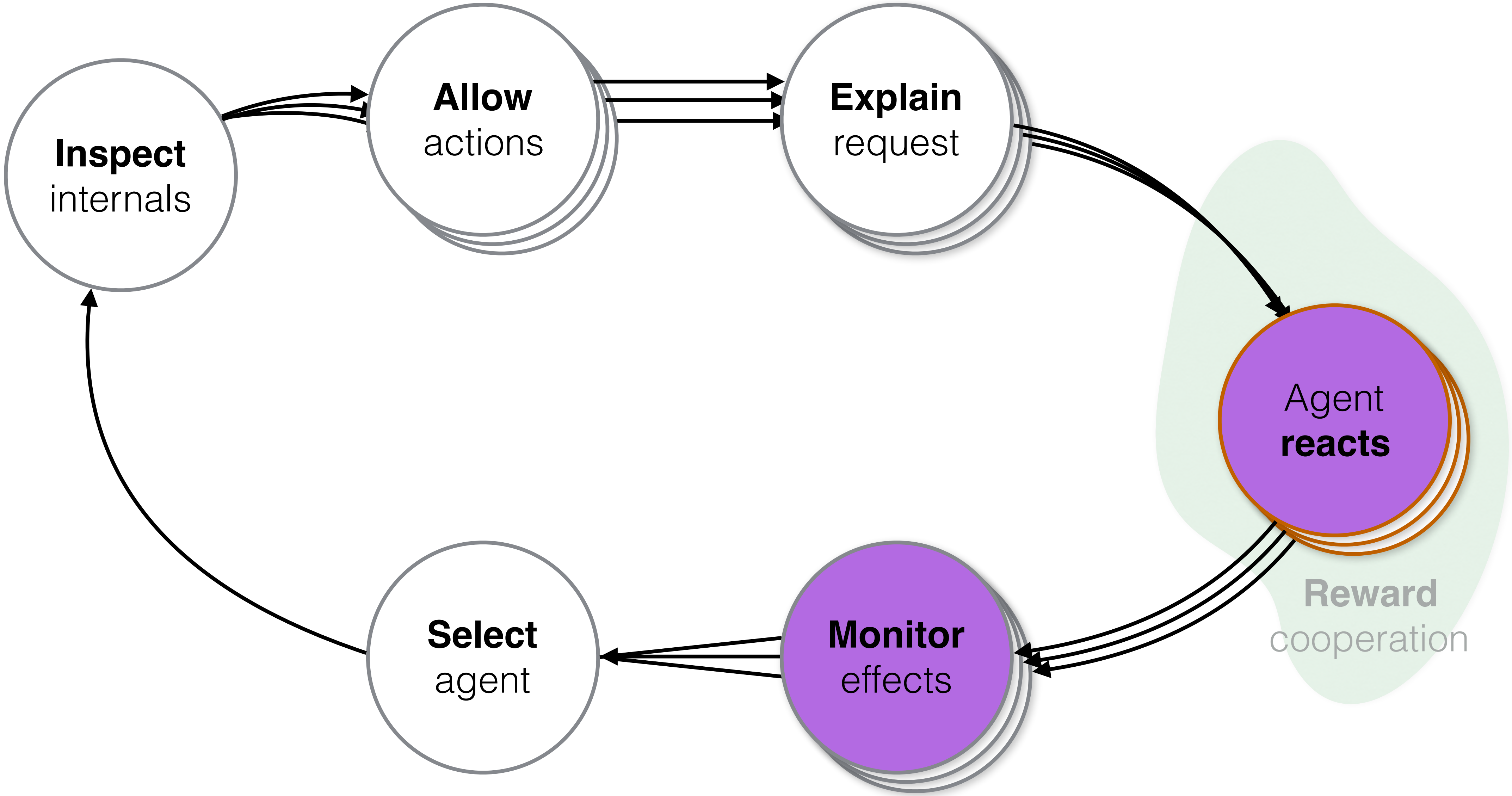
Internal Software Development

API Design



Internal Software Development

Testing. Bug reports.



The Elements of Decision Alignment

	Human to Human	Human to/from Object	Object to Object
Select agent	Trademark Chain of custody	App stores White and black lists	Trusted developer Same origin
Inspect internals	Accounting controls	Trusted path URL bar	Types, Verification Open source eyeballs
Allow actions	Law, Contracts	App permissions Powerbox	Security Protection patterns
Explain request	Language	User interface	Abstraction
Reward cooperation	Economics Incentive Alignment	Objective functions	Machine learning Agorics
Monitor effects	Reviews, Complaints Word of mouth	Bug reports	Contracts, Testing Backprop

Tuning Tradeoffs

Select agent

Inspect internals

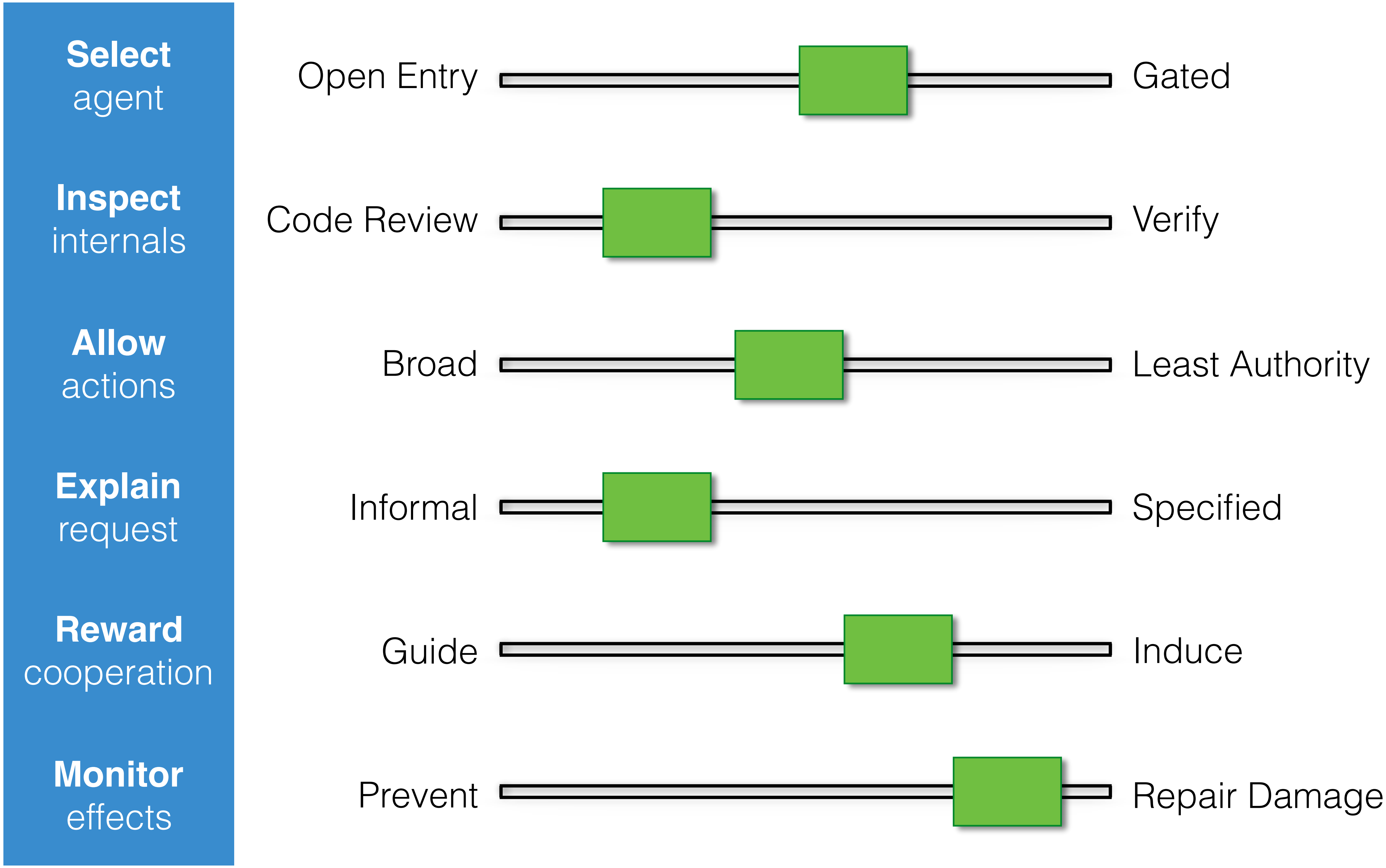
Allow actions

Explain request

Reward cooperation

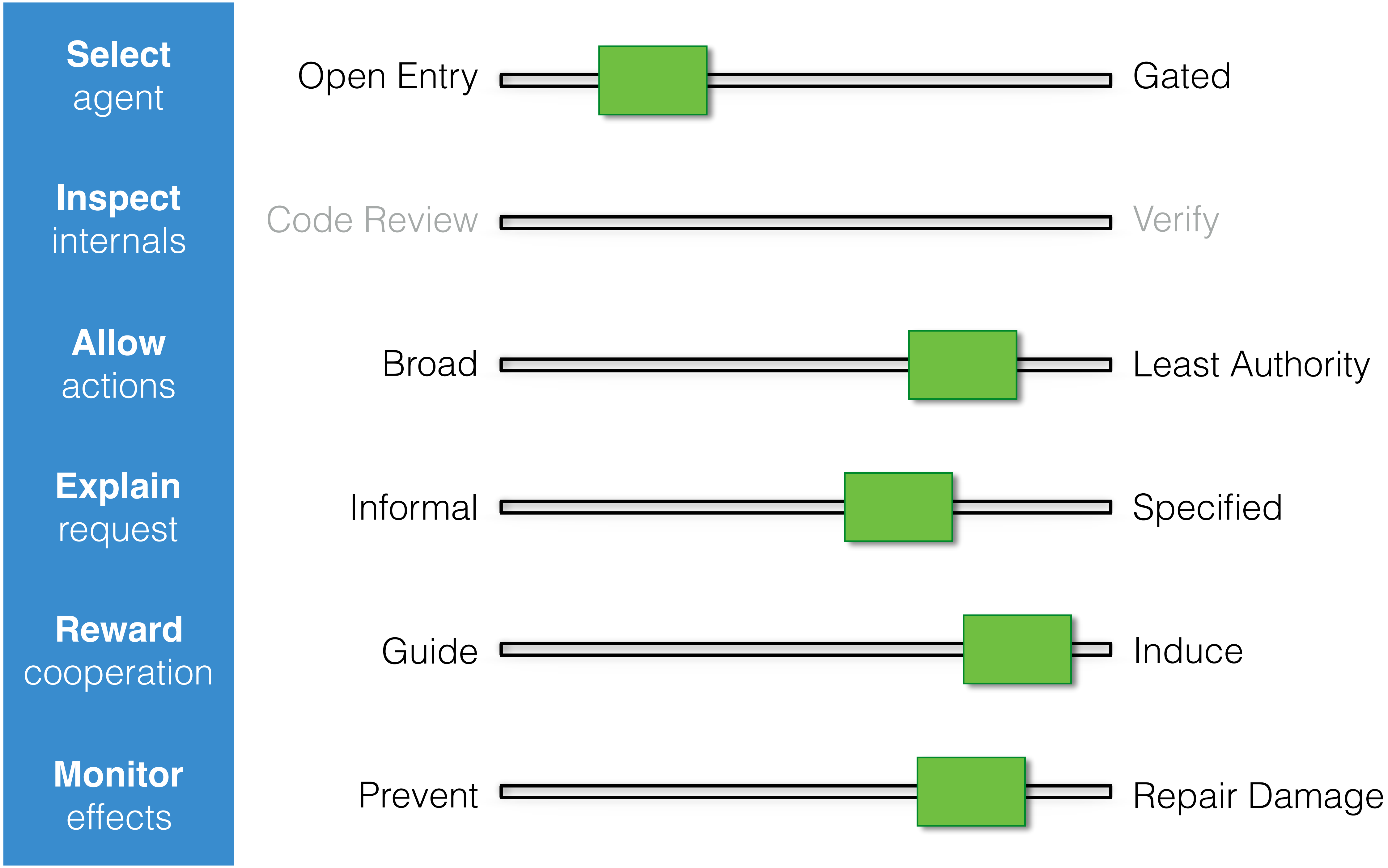
Monitor effects

Tuning Tradeoffs



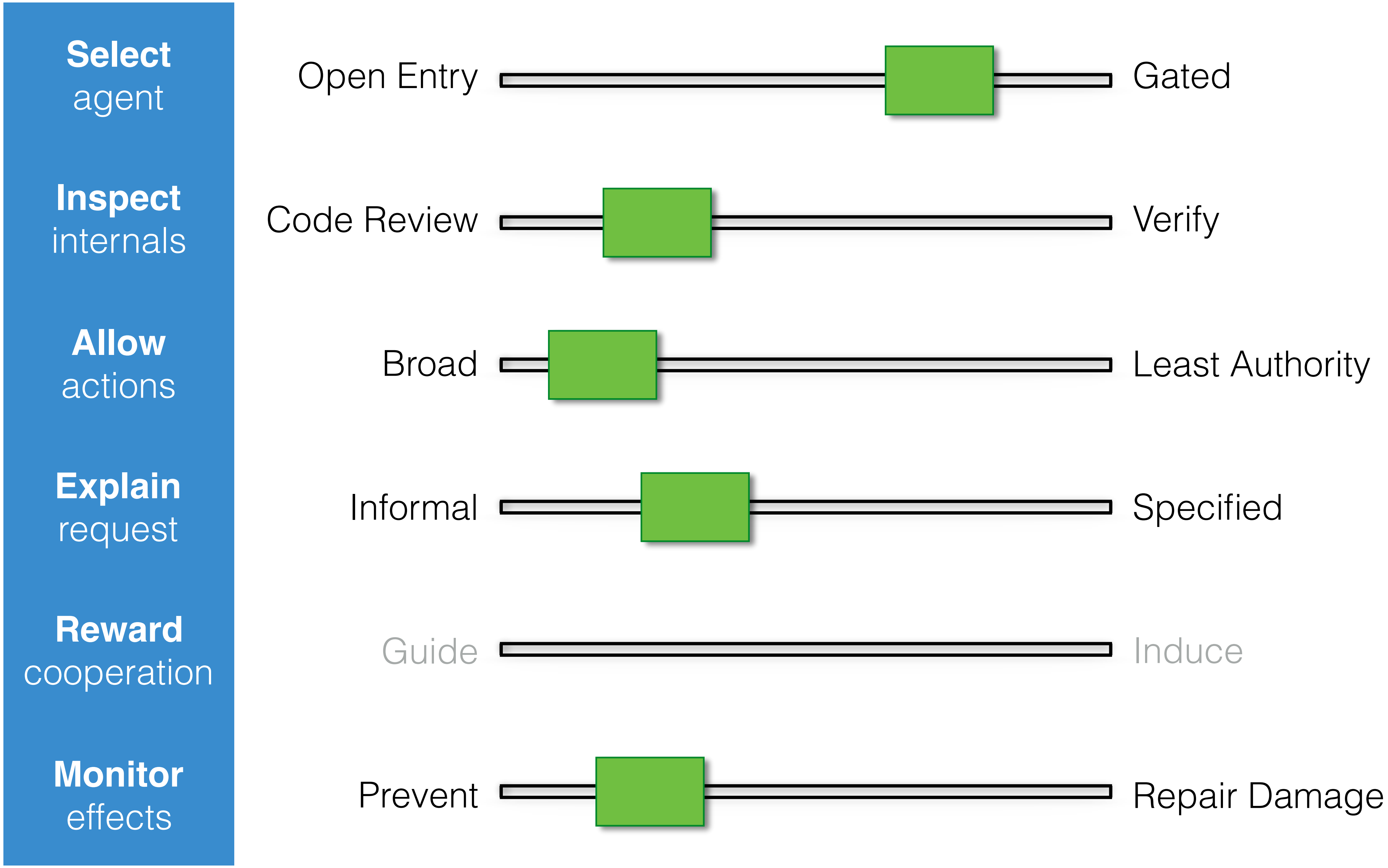
Select agent	Open Entry	Gated
Inspect internals	Code Review	Verify
Allow actions	Broad	Least Authority
Explain request	Informal	Specified
Reward cooperation	Guide	Induce
Monitor effects	Prevent	Repair Damage

Package Delivery Business



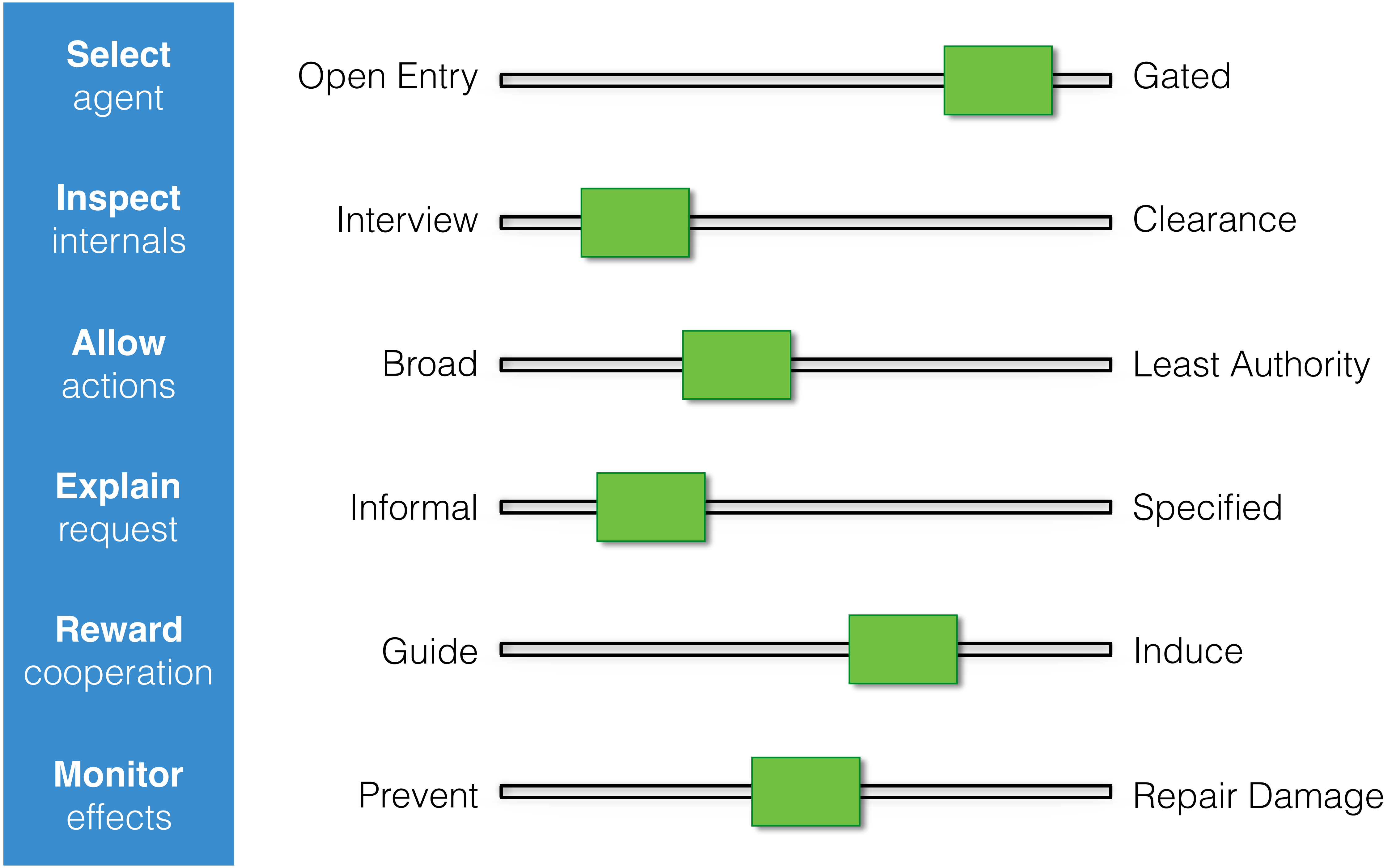
Select agent	Open Entry	Gated
Inspect internals	Code Review	Verify
Allow actions	Broad	Least Authority
Explain request	Informal	Specified
Reward cooperation	Guide	Induce
Monitor effects	Prevent	Repair Damage

Internal Software Development



Select agent	Open Entry	Gated
Inspect internals	Code Review	Verify
Allow actions	Broad	Least Authority
Explain request	Informal	Specified
Reward cooperation	Guide	Induce
Monitor effects	Prevent	Repair Damage

Organizational Employment



Select agent — Open Entry / Gated

Inspect internals — Interview / Clearance

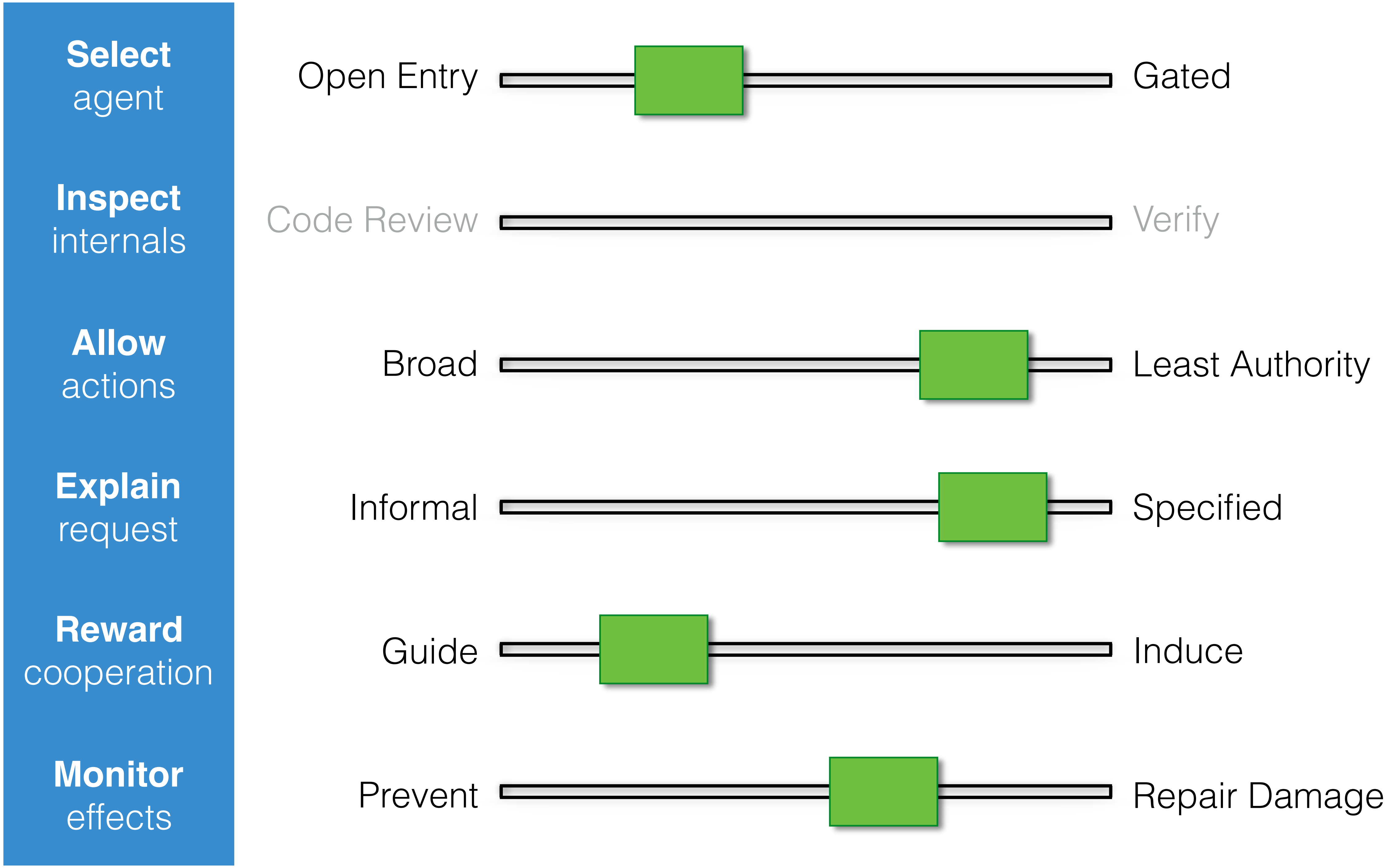
Allow actions — Broad / Least Authority

Explain request — Informal / Specified

Reward cooperation — Guide / Induce

Monitor effects — Prevent / Repair Damage

Safe Plugin Boundary



Select agent

Open Entry … Gated

Inspect internals

Code Review … Verify

Allow actions

Broad … Least Authority

Explain request

Informal … Specified

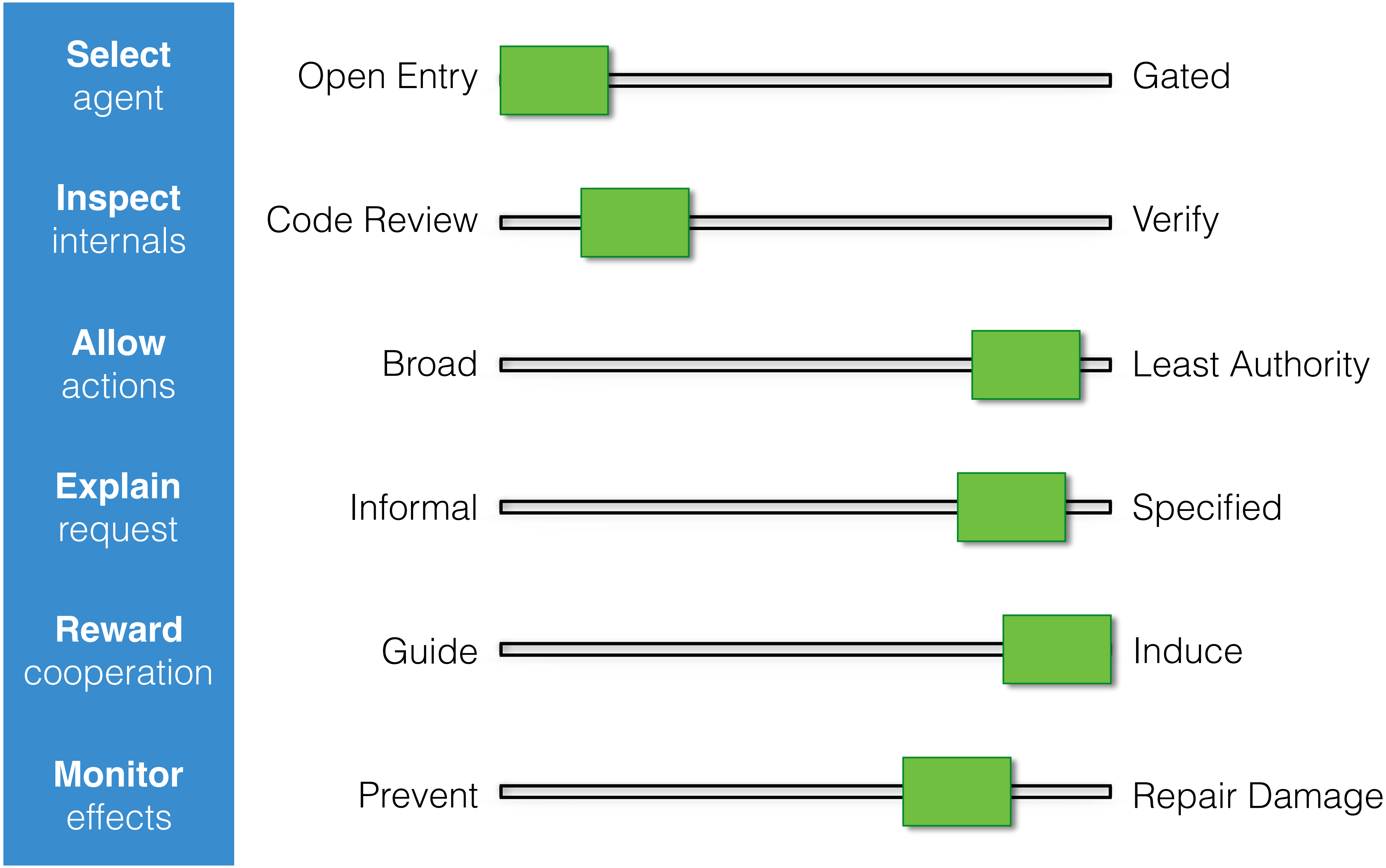
Reward cooperation

Guide … Induce

Monitor effects

Prevent … Repair Damage

Bitcoin, Etherium



Select agent

Open Entry — Gated

Inspect internals

Code Review — Verify

Allow actions

Broad — Least Authority

Explain request

Informal — Specified

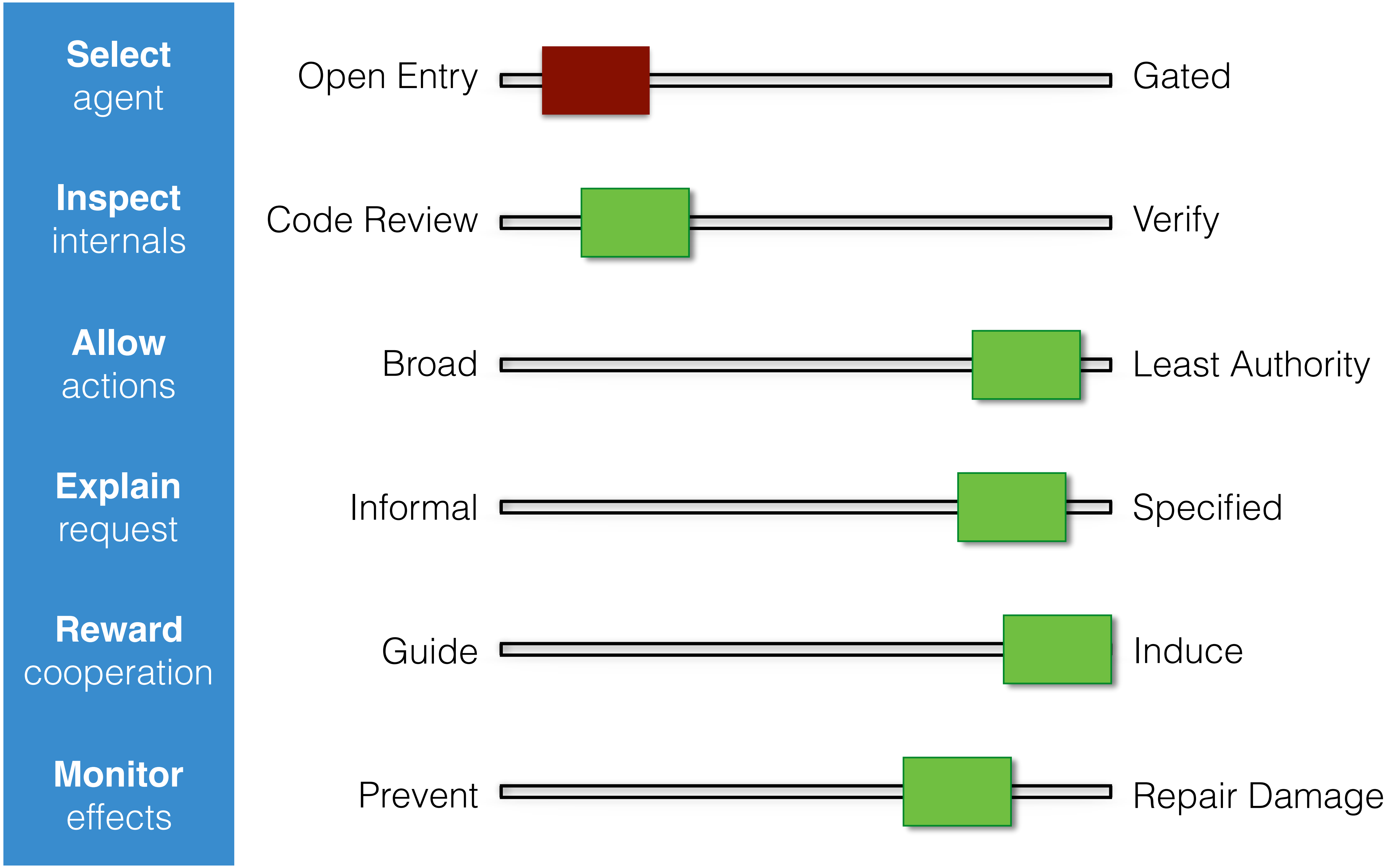
Reward cooperation

Guide — Induce

Monitor effects

Prevent — Repair Damage

Etherium Hard Fork



Select agent

Open Entry — Gated

Inspect internals

Code Review — Verify

Allow actions

Broad — Least Authority

Explain request

Informal — Specified

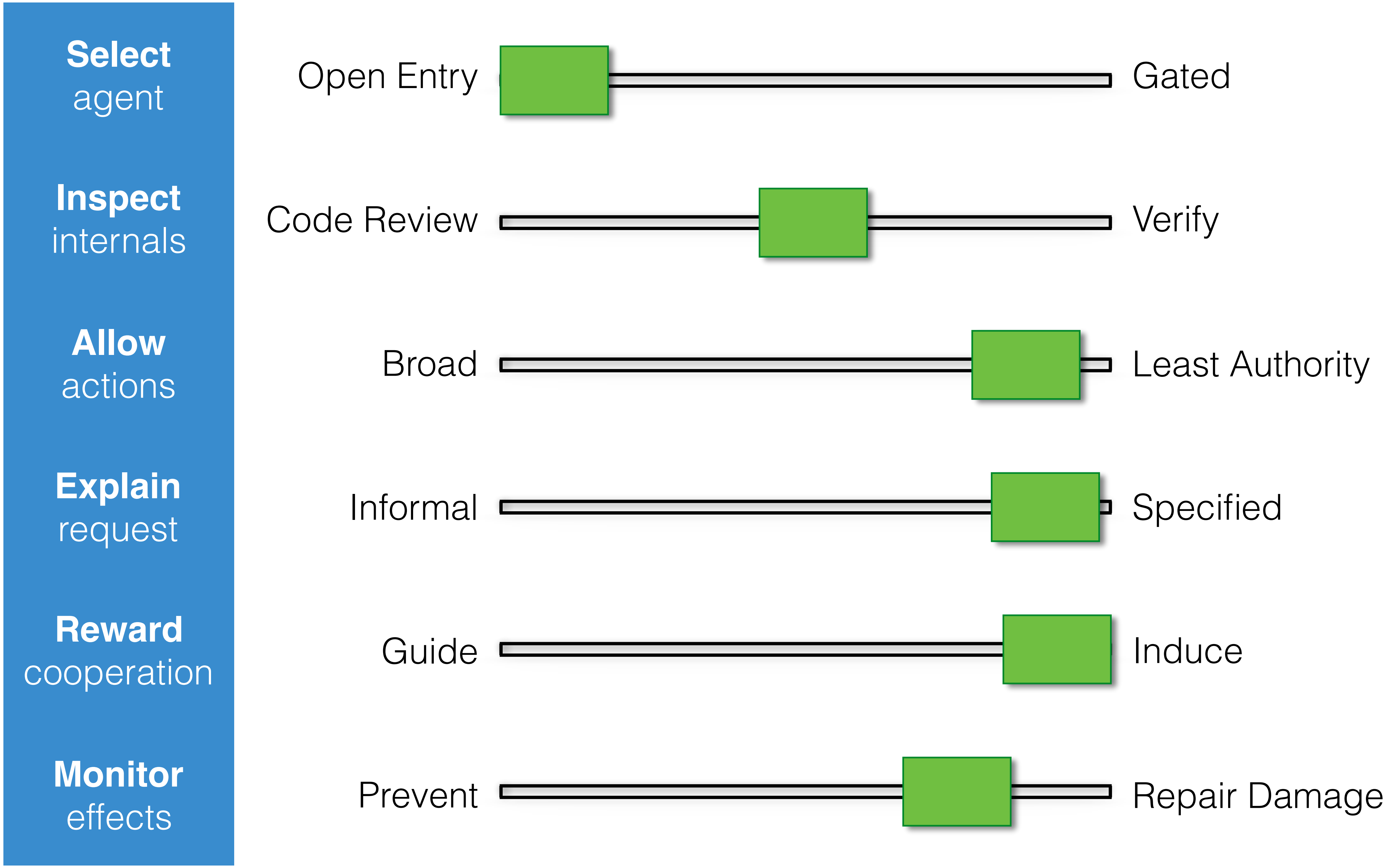
Reward cooperation

Guide — Induce

Monitor effects

Prevent — Repair Damage

Etherium once repaired



Select agent — Open Entry / Gated

Inspect internals — Code Review / Verify

Allow actions — Broad / Least Authority

Explain request — Informal / Specified

Reward cooperation — Guide / Induce

Monitor effects — Prevent / Repair Damage

Building Reliable Voting Machine Software

Ka-Ping Yee, 2007 dissertation

Programmer may wish to bias the election.

Programmer and code as untrusted agent.

Building Reliable Voting Machine Software

Ka-Ping Yee, 2007 dissertation

Programmer may wish to bias the election.

Must write simple code that seems obviously correct.

400 lines of simple code in simple language.
Extensive rationale justifying each line.

Building Reliable Voting Machine Software

Ka-Ping Yee, 2007 dissertation

Programmer may wish to bias the election.

Must write simple code that seems obviously correct.

Subject to extremely intense review.

Intense review of simple code is effective
at spotting *accidental* bugs and vulnerabilities.

Building Reliable Voting Machine Software

Ka-Ping Yee, 2007 dissertation

Programmer may wish to bias the election.

Must write simple code that seems obviously correct.

Subject to extremely intense review.

I am one of the reviewers who failed to find malicious bugs.
None succeeded at finding all three bugs.

Building Reliable Voting Machine Software

Ka-Ping Yee, 2007 dissertation

Programmer may wish to bias the election.

Must write simple code that seems obviously correct.

Subject to extremely intense review.

Malicious bugs easily evade detection by review *or* testing.

Building Reliable Voting Machine Software

Ka-Ping Yee, 2007 dissertation

Programmer may wish to bias the election.

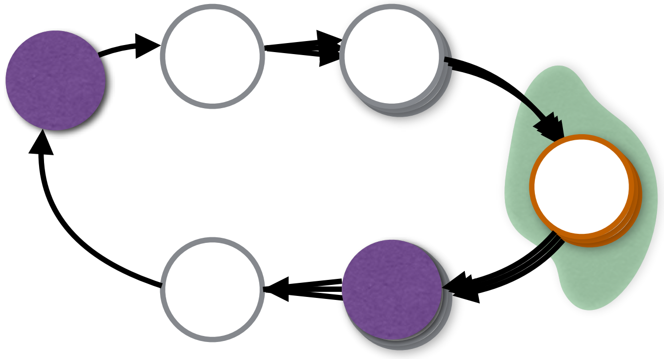
Must write simple code that seems obviously correct.

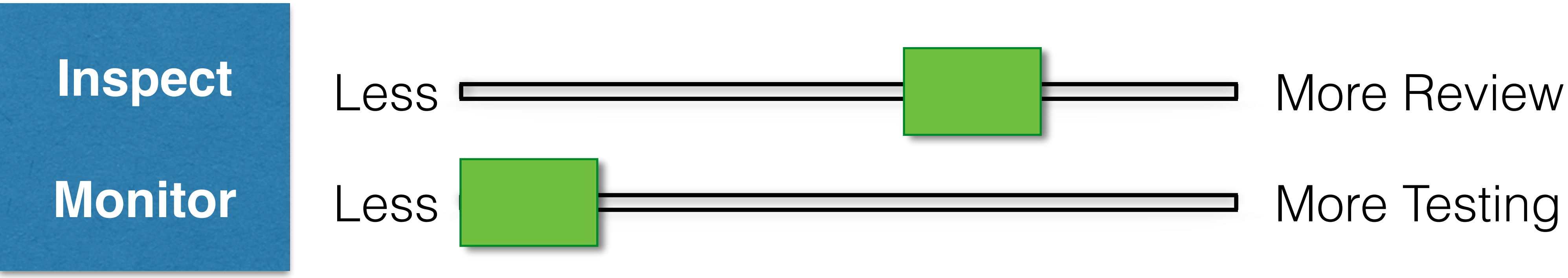
Subject to extremely intense review.

Malicious bugs easily evade detection by review *or* testing.

Harder to evade detection by review *and* testing.

Inspect + Monitor





Inspect

Less — More Review

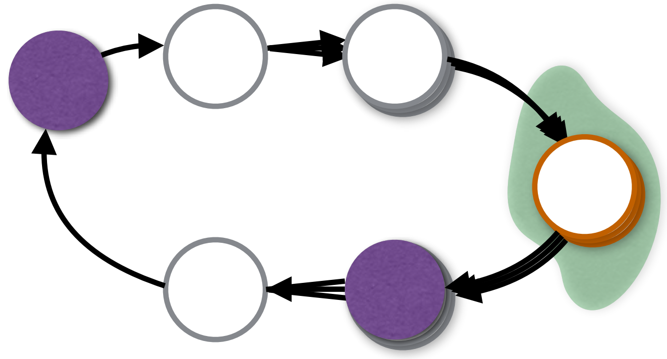
Monitor

Less — More Testing

```
for (i = 0; i <= limit; i++)
```

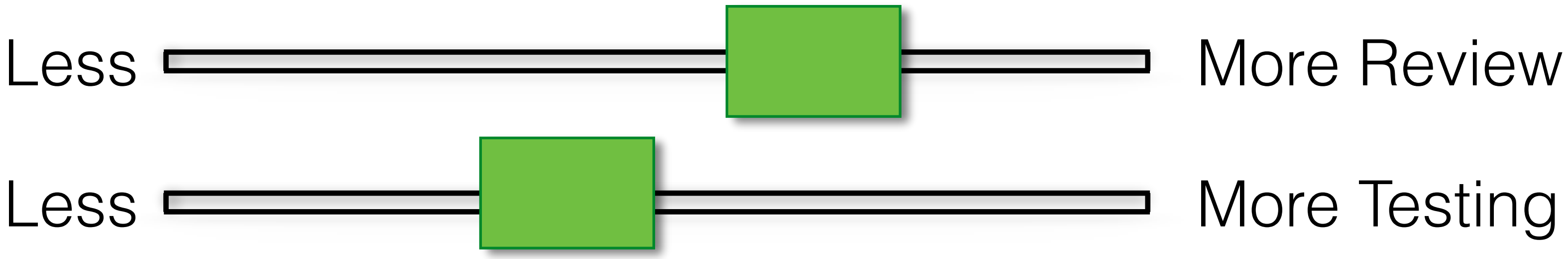
Looks fine.

Inspect + Monitor



Inspect

Monitor



Less

More Review

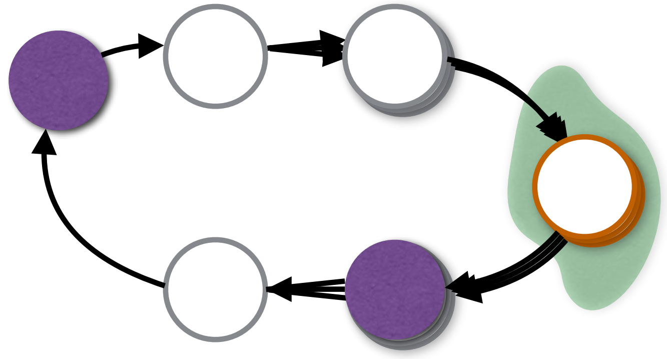
Less

More Testing

```
for (i = 0; i <= limit; i++)
```

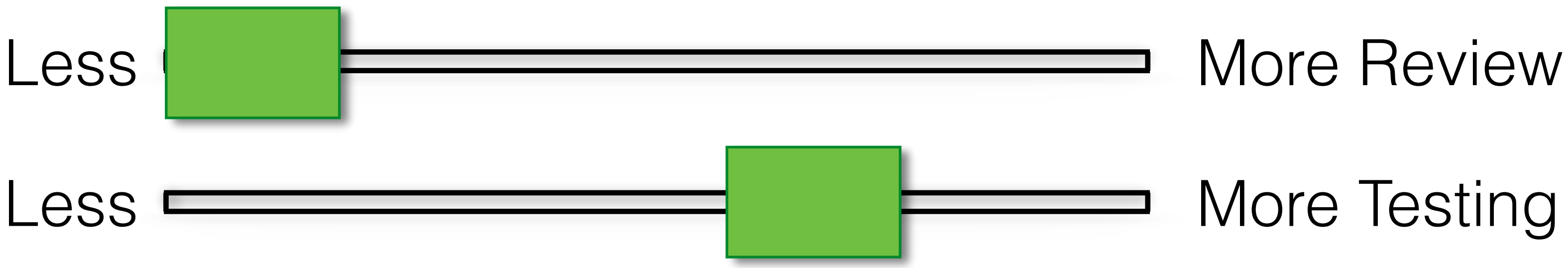
Looks fine.
Fails on zero and one.

Inspect + Monitor



Inspect

Monitor



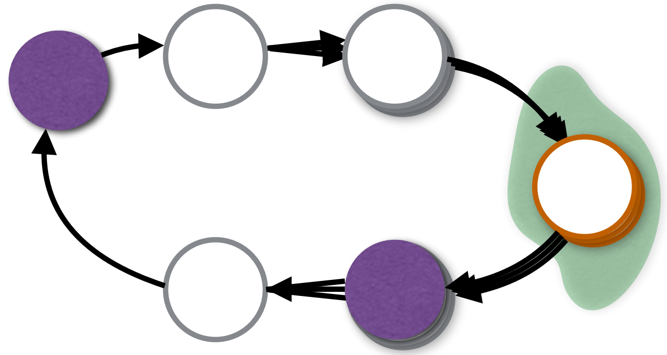
```
for (i = 0; i <= limit; i++)
```

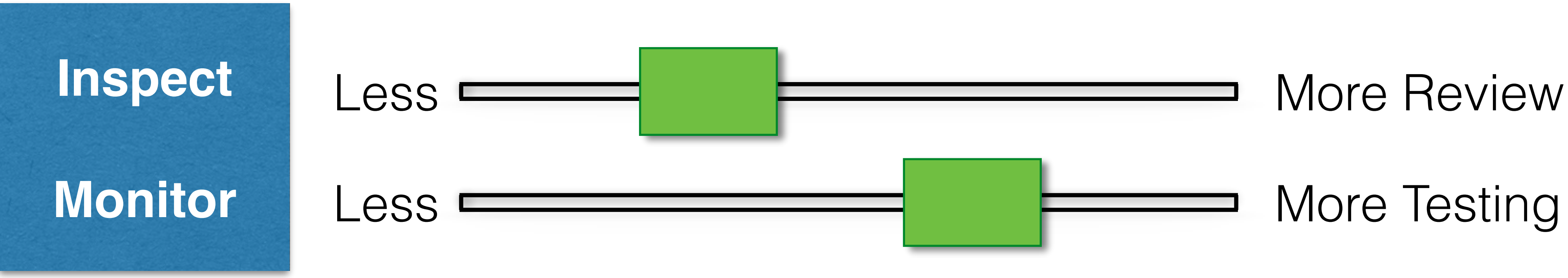
Looks fine.
Fails on zero and one.

```
for (i = 0; i < limit; i++)
  if (j === 72374928)
```

Passes all tests.

Inspect + Monitor





Inspect

Monitor

Less More Review

Less More Testing

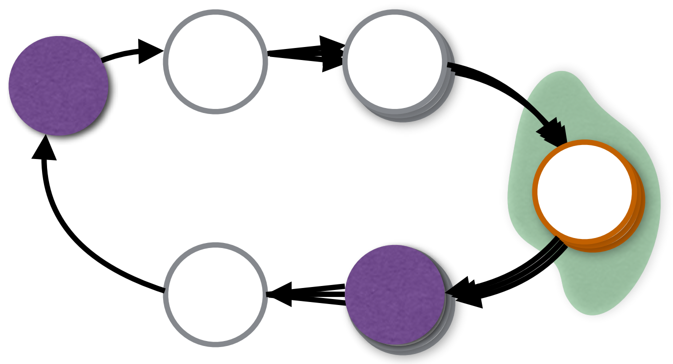
```
for (i = 0; i <= limit; i++)
```

Looks fine.
Fails on zero and one.

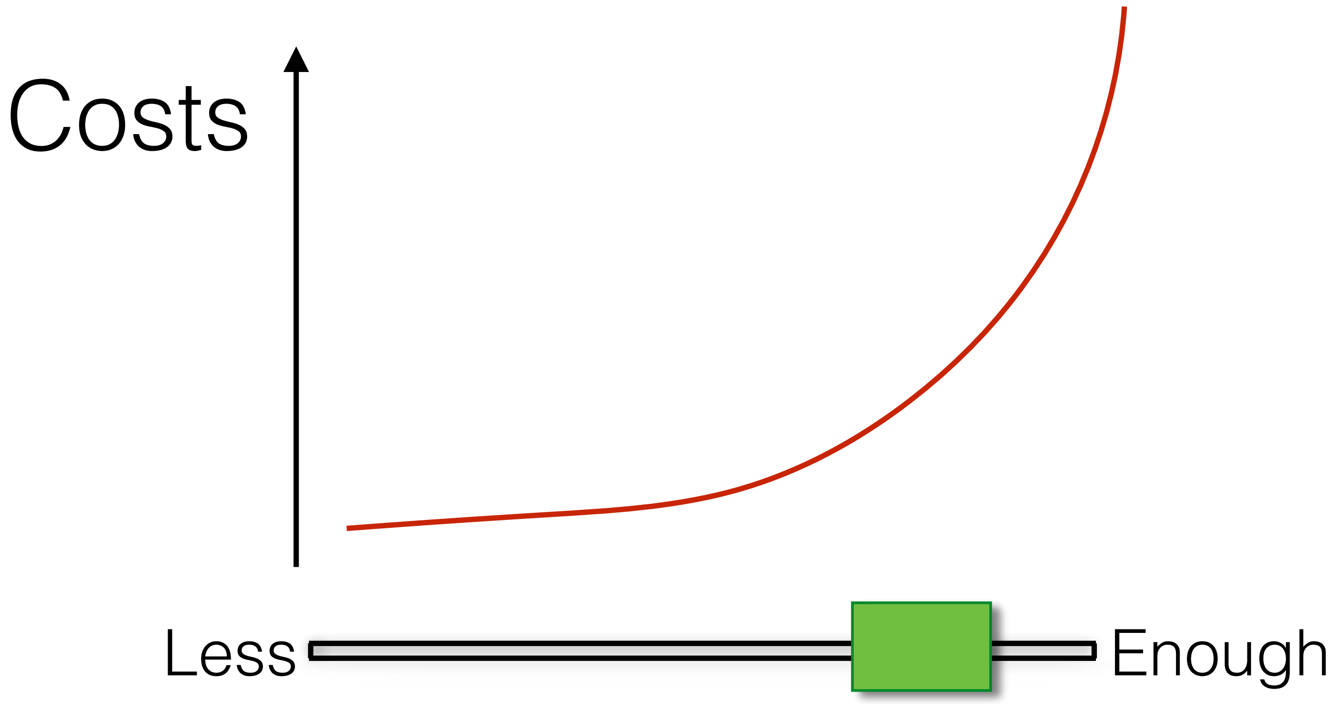
```
for (i = 0; i < limit; i++)
  if (j === 72374928)
```

Looks weird.
Passes all tests.

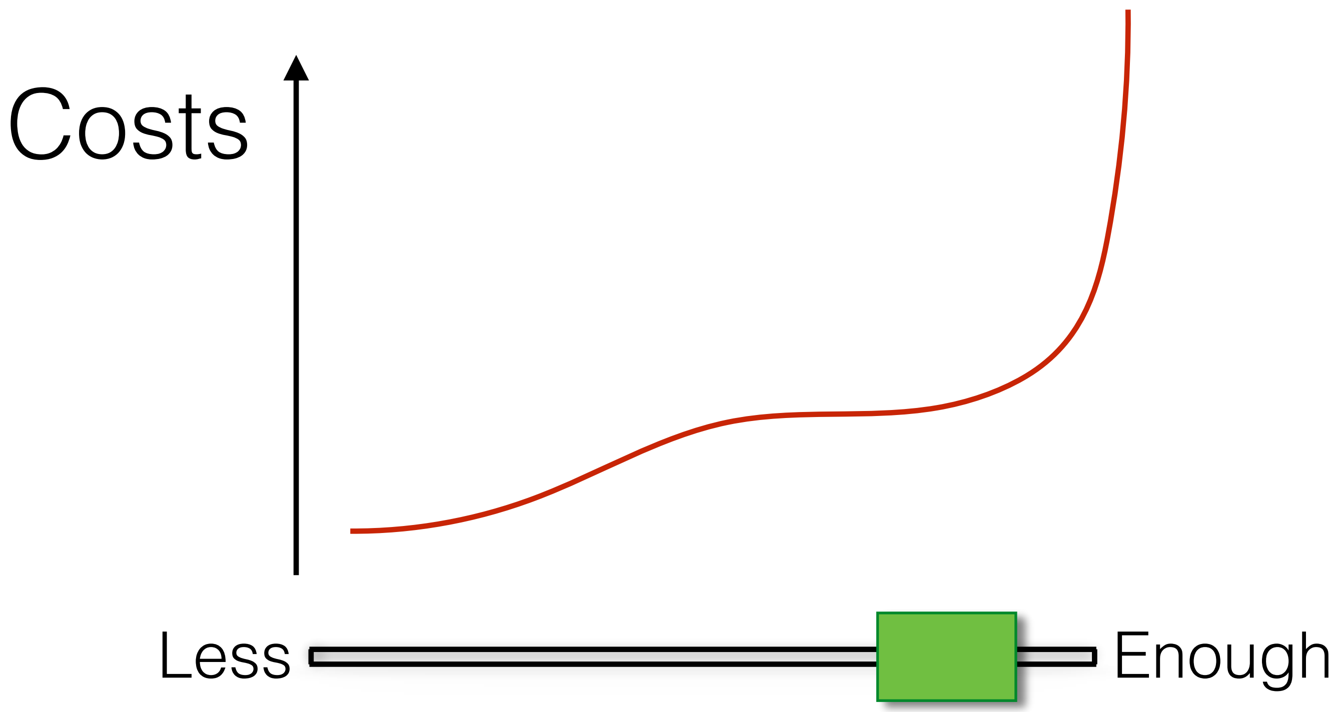
Compose Compromises



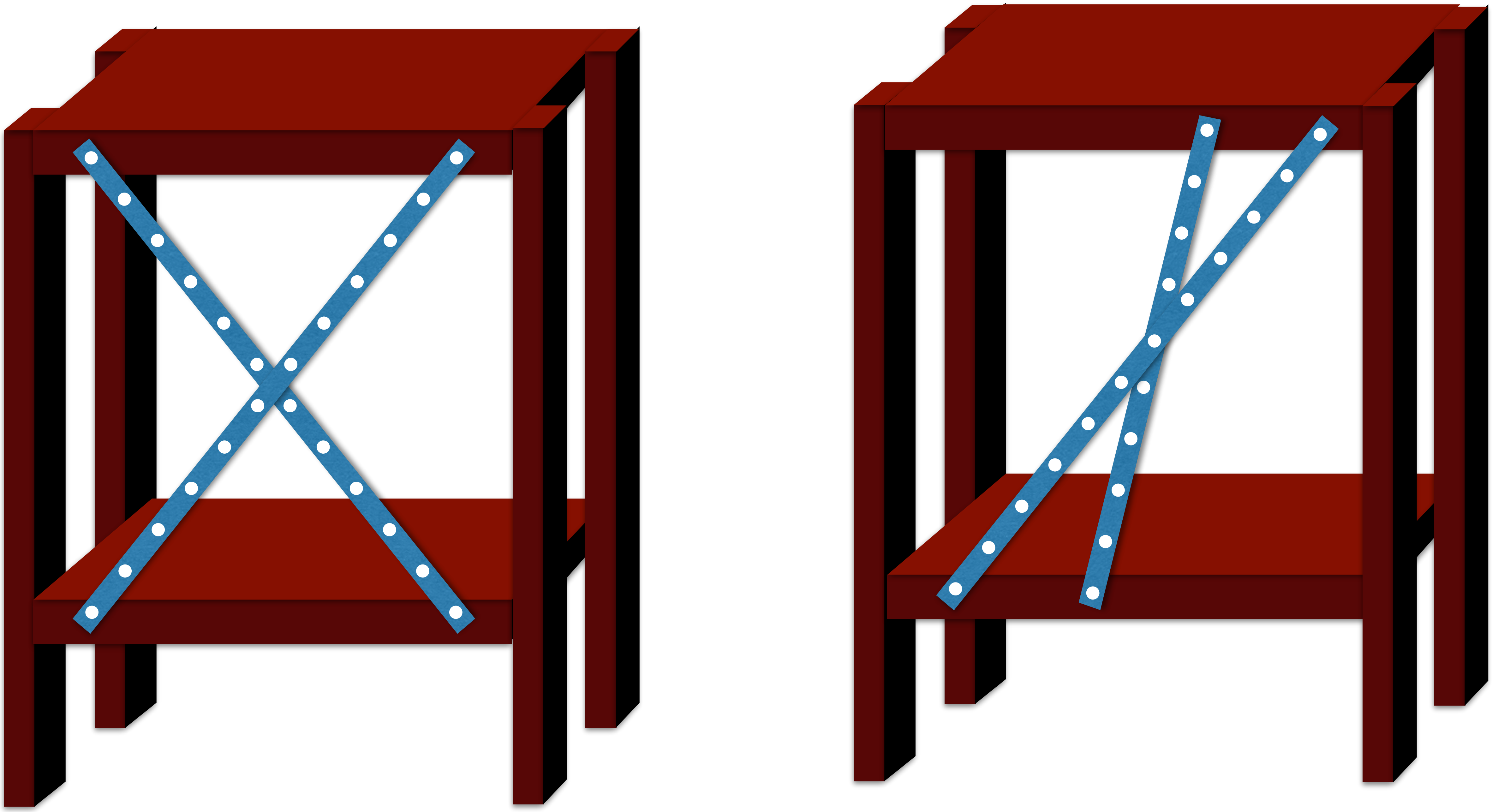
Inspect



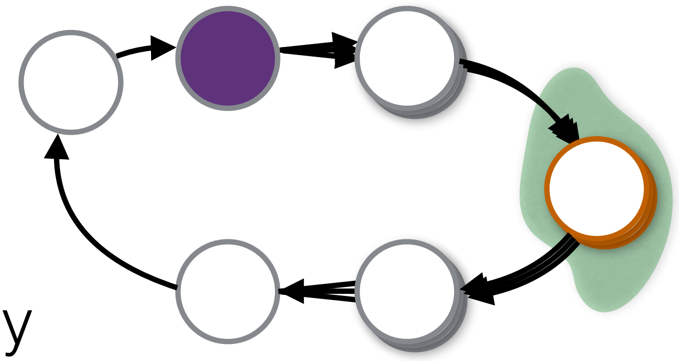
Monitor



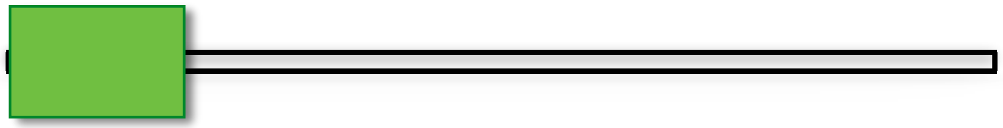
Cross Bracing



Package Delivery

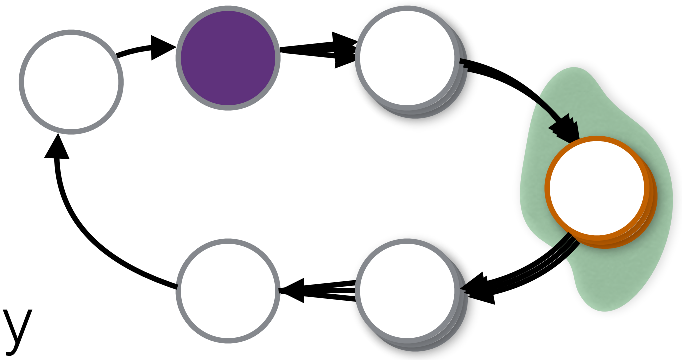


Allow

Broad  Least Authority

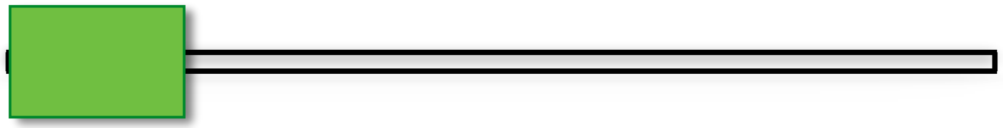
Broad Authority

Package Delivery

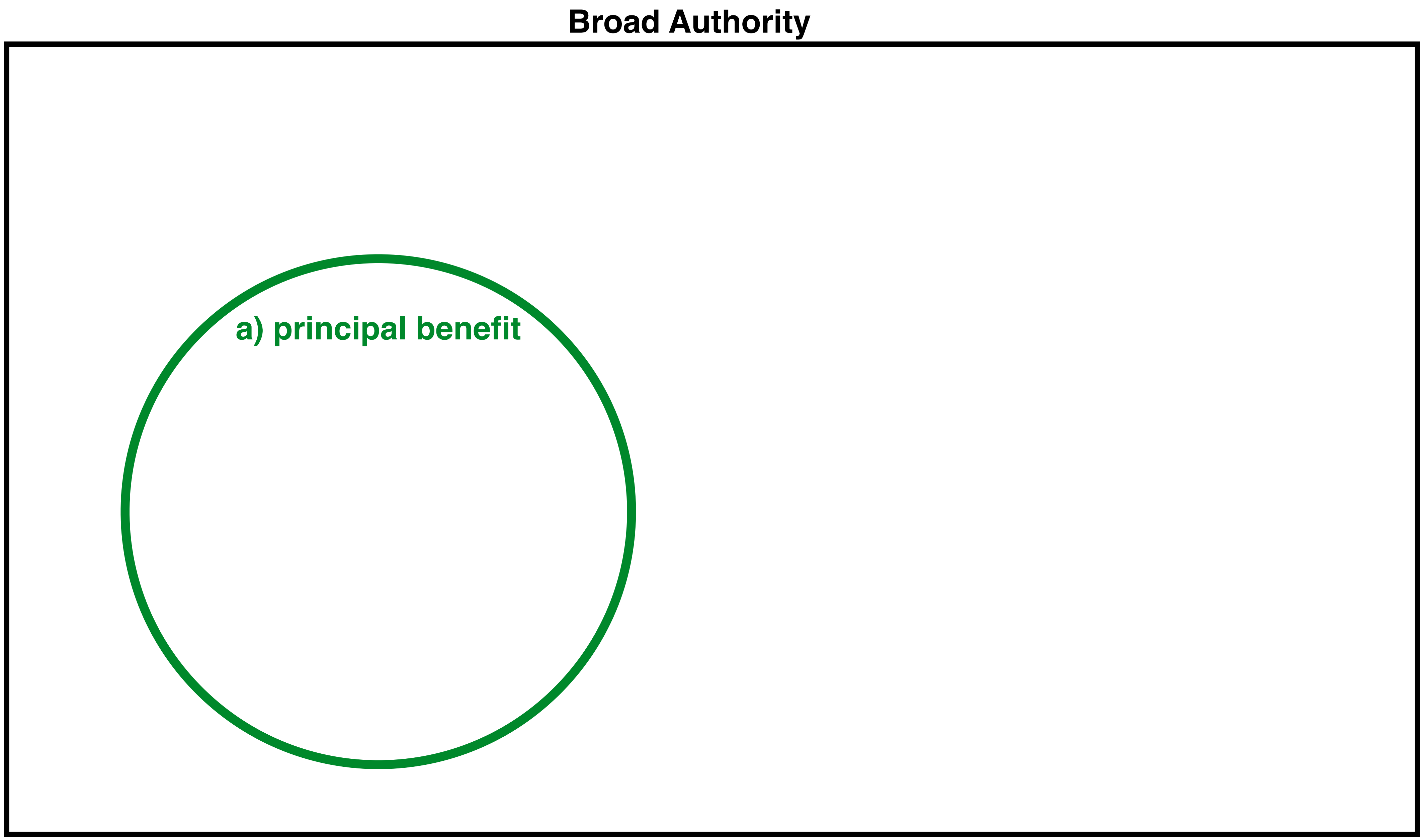


Allow

Broad

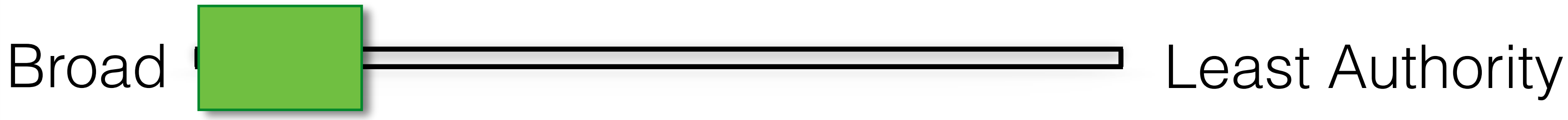


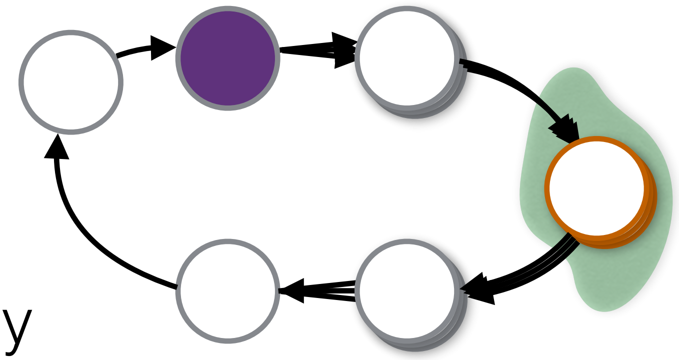
Least Authority

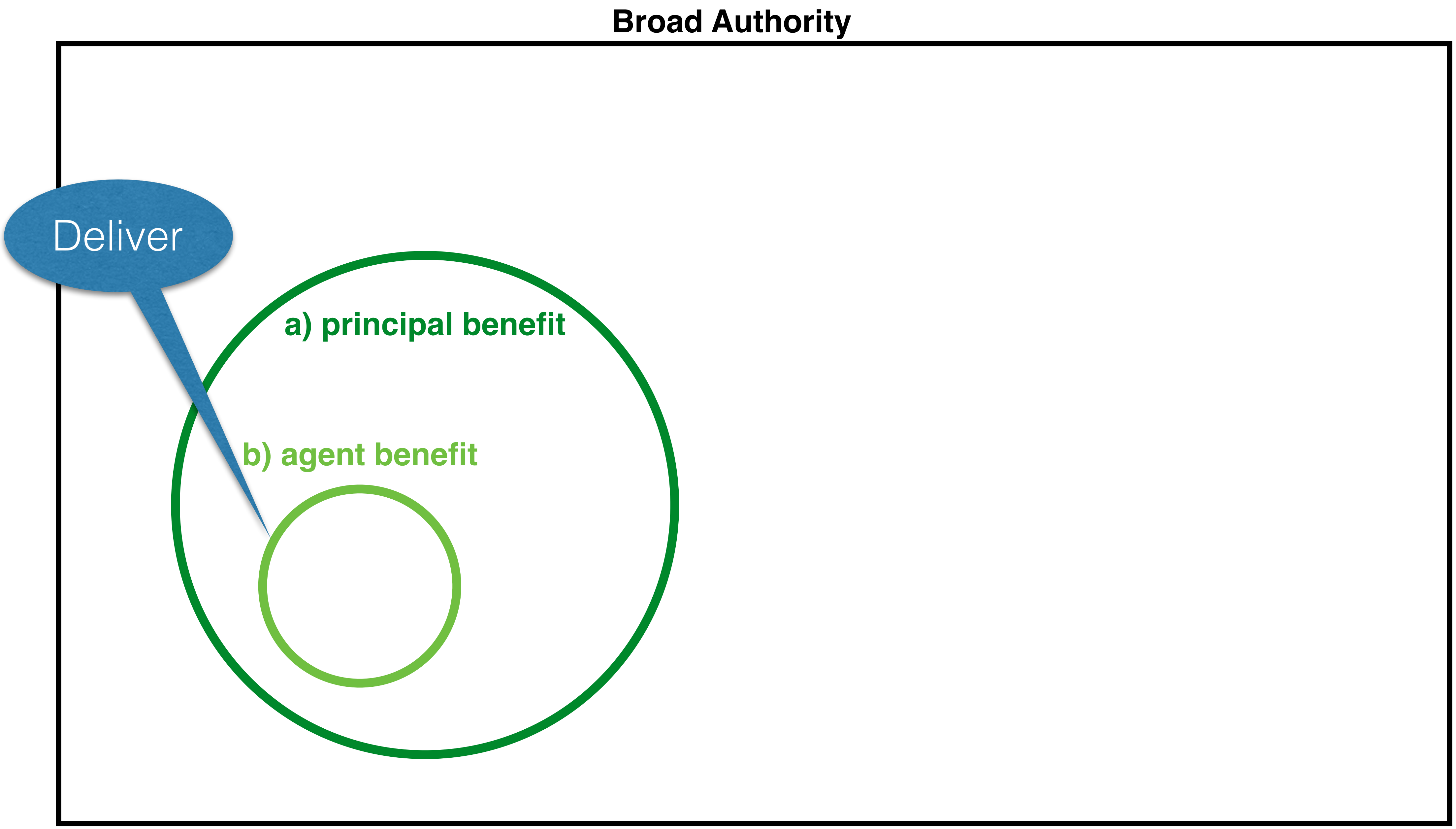


Package Delivery

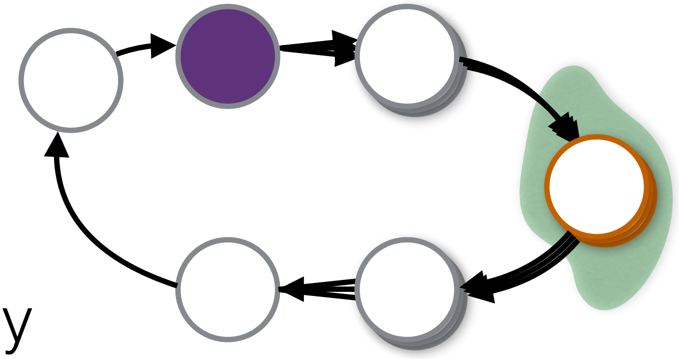
Allow







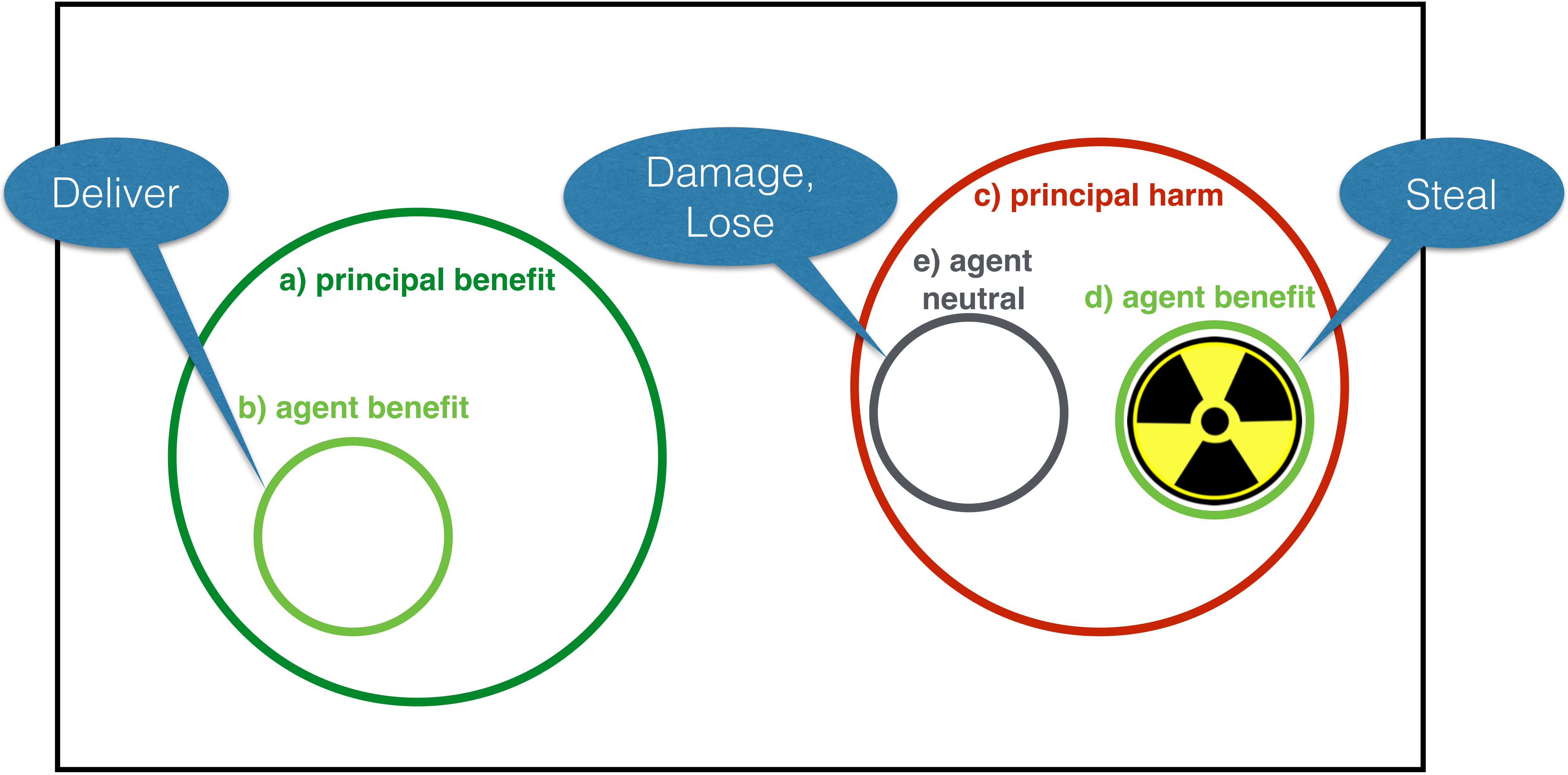
Package Delivery



Allow

Broad 🟩 Least Authority

Broad Authority



Deliver

a) principal benefit

b) agent benefit

Damage, Lose

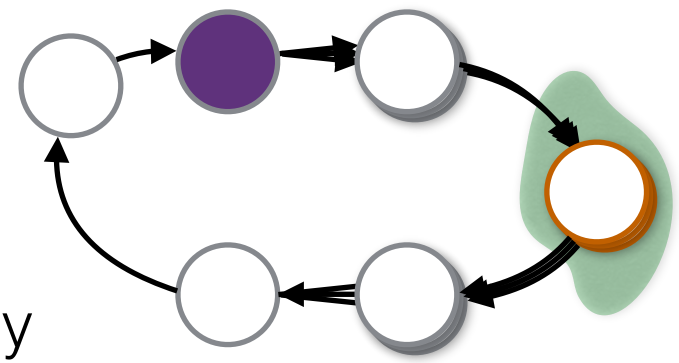
c) principal harm

e) agent neutral

d) agent benefit

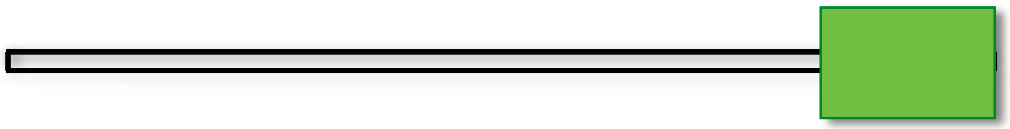
Steal

Package Delivery

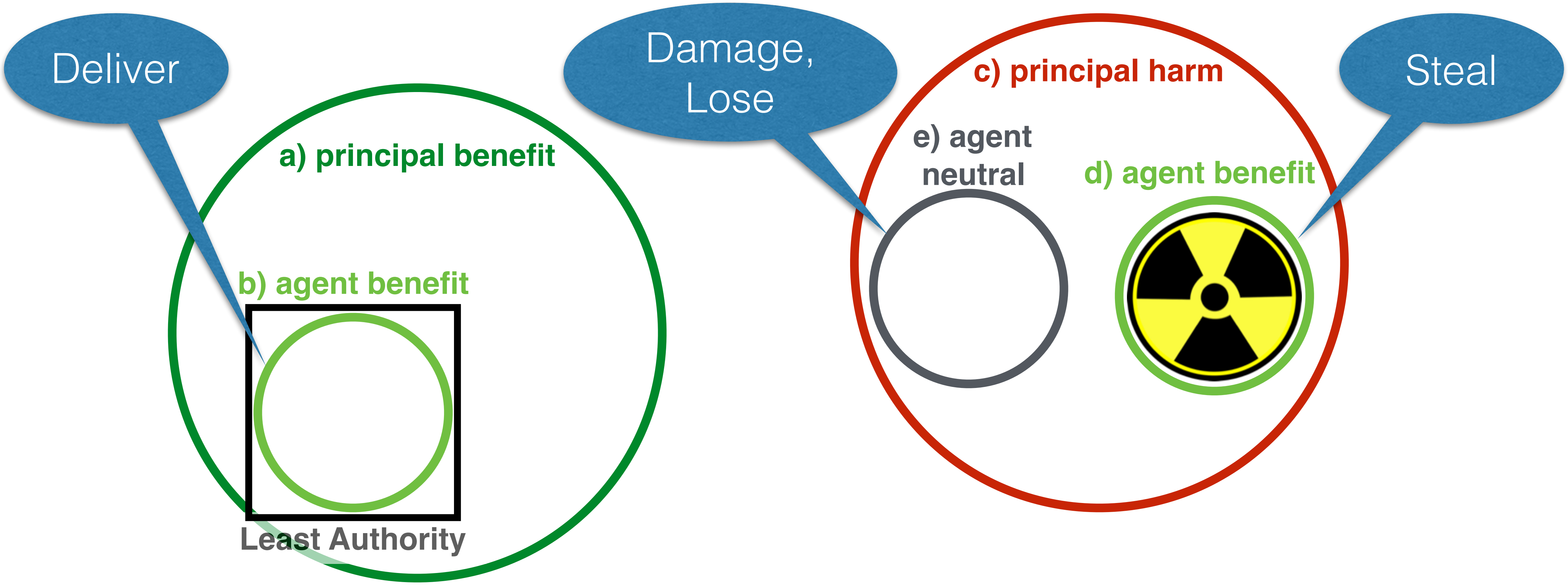


Allow

Broad



Least Authority



Deliver

Damage, Lose

Steal

a) principal benefit

b) agent benefit

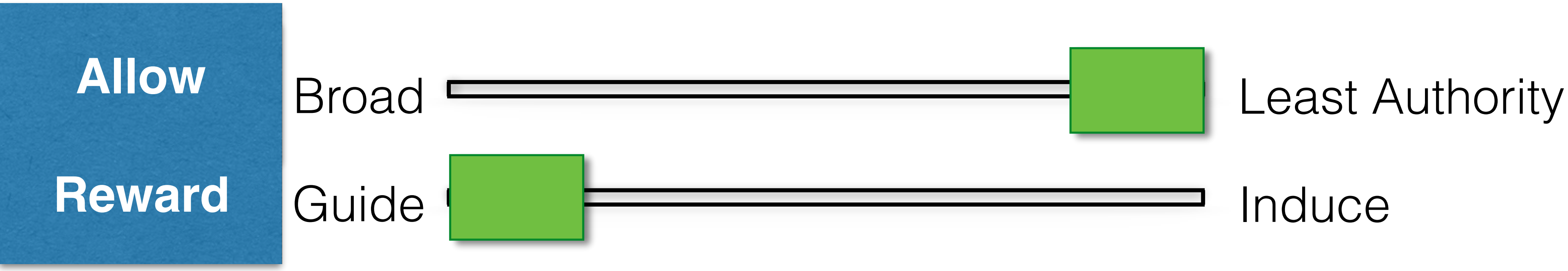
Least Authority

c) principal harm

e) agent neutral

d) agent benefit

Allow + Reward



Allow

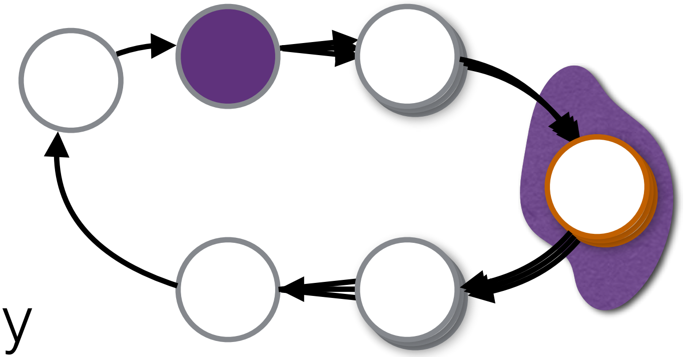
Reward

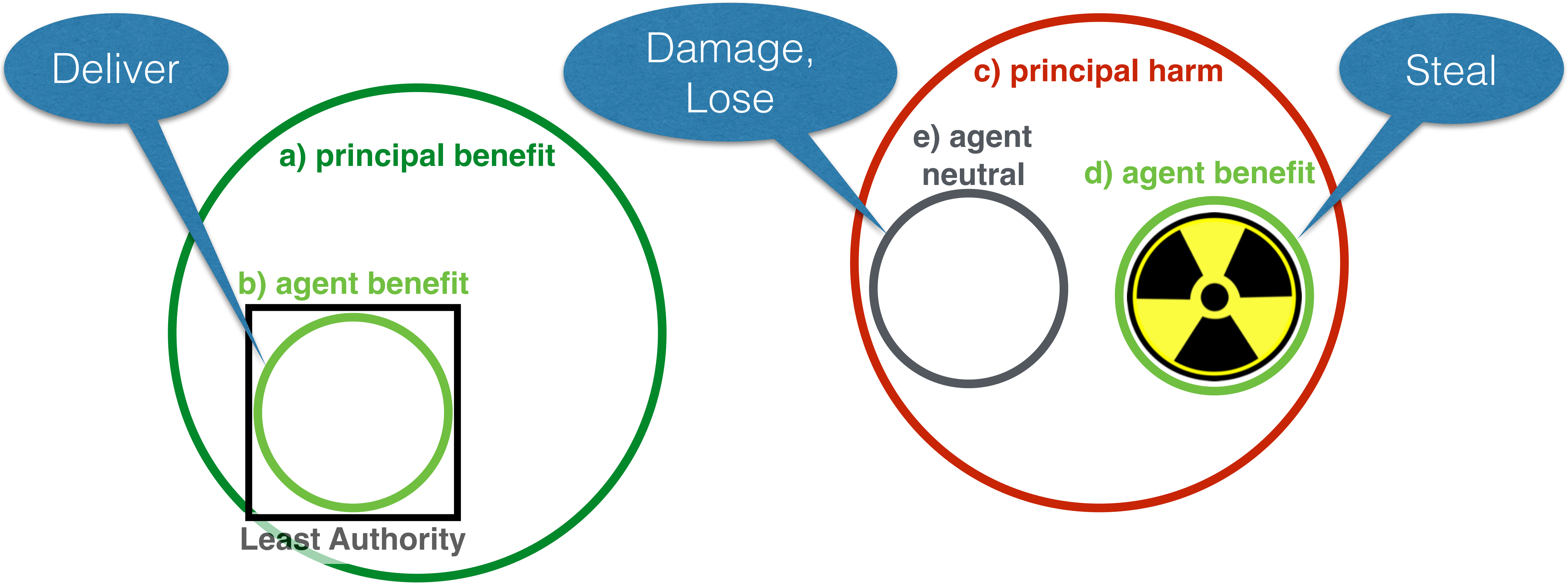
Broad

Least Authority

Guide

Induce





Deliver

a) principal benefit

b) agent benefit

Least Authority

Damage, Lose

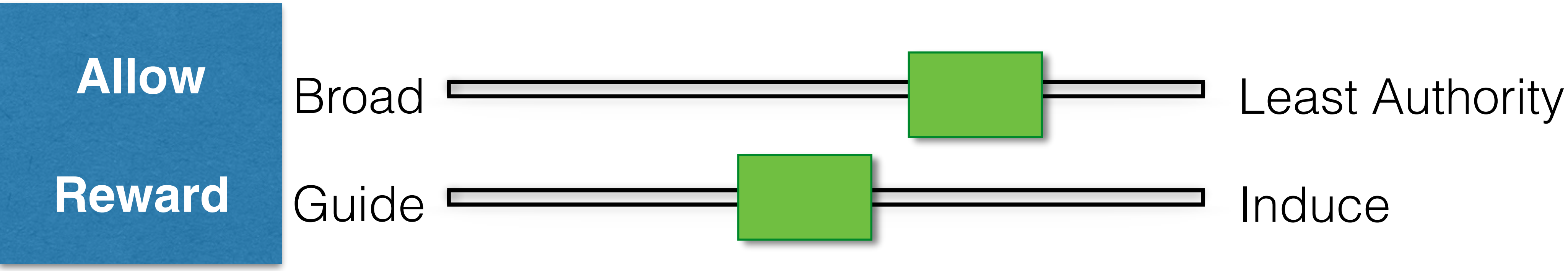
c) principal harm

e) agent neutral

d) agent benefit

Steal

Allow + Reward



Allow

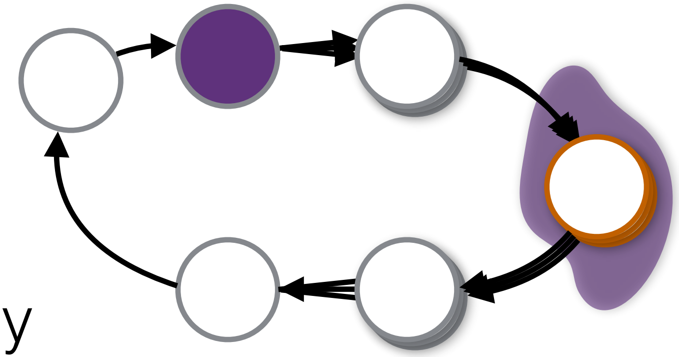
Reward

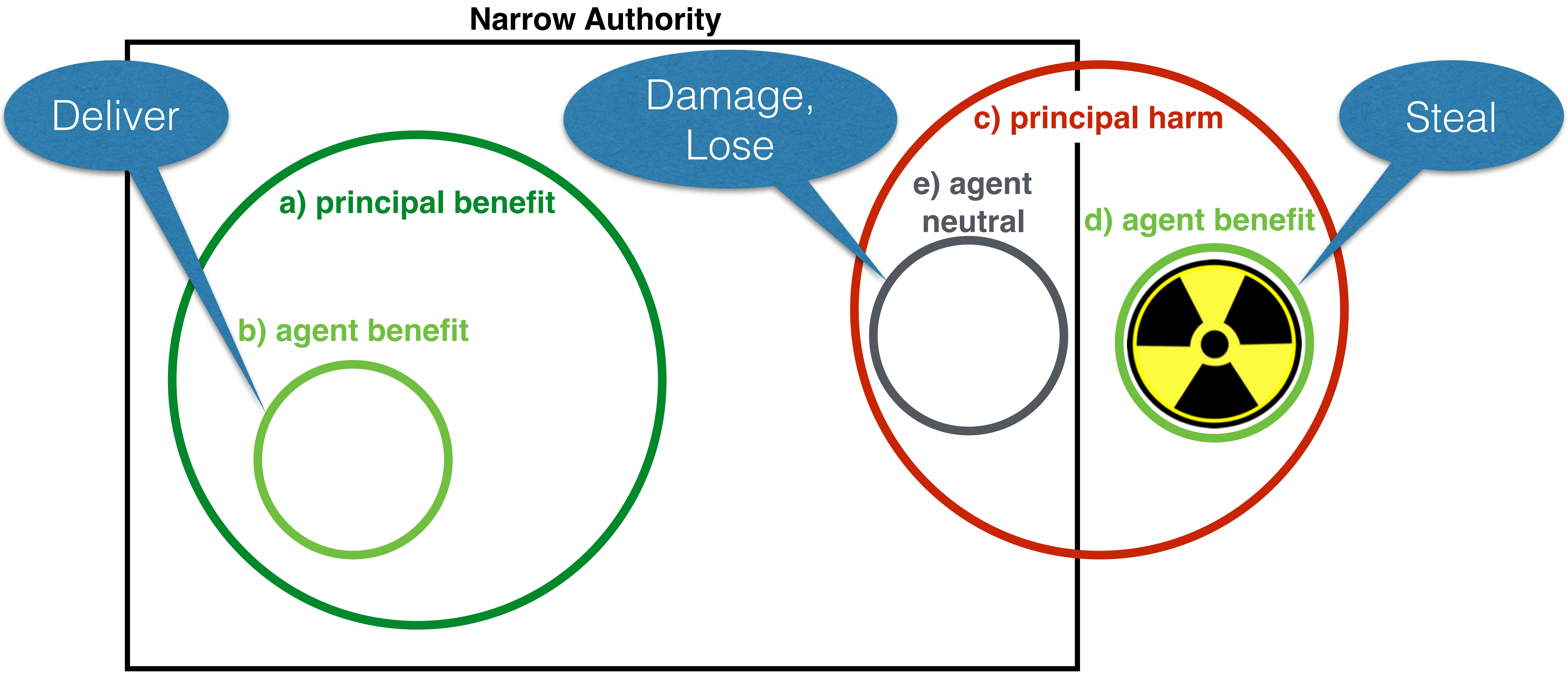
Broad

Least Authority

Guide

Induce





Narrow Authority

Deliver

Damage, Lose

Steal

a) principal benefit

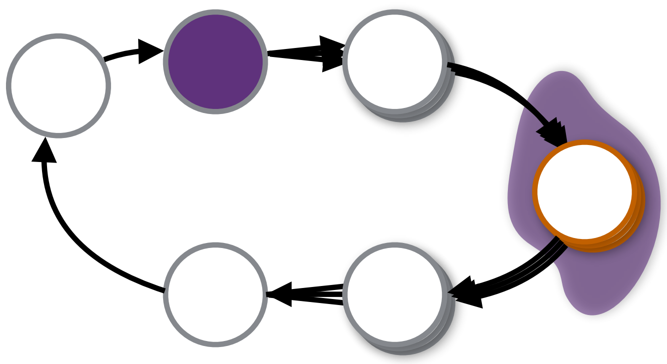
b) agent benefit

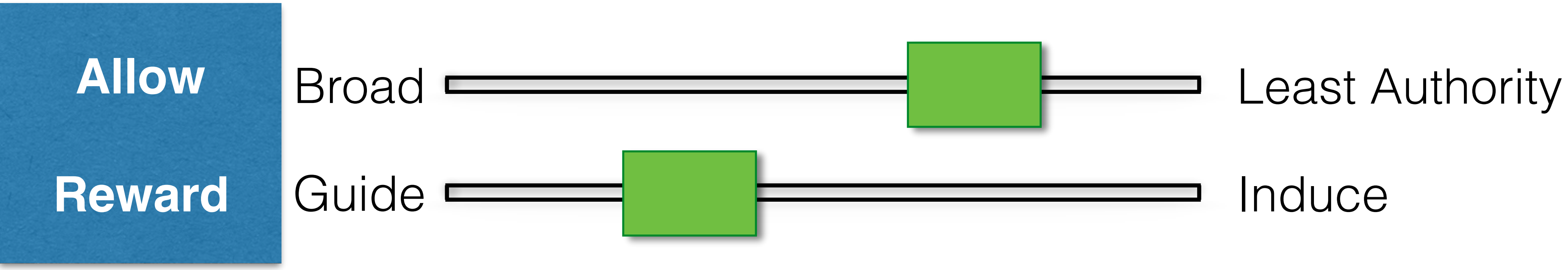
c) principal harm

e) agent neutral

d) agent benefit

Allow + Reward





Allow

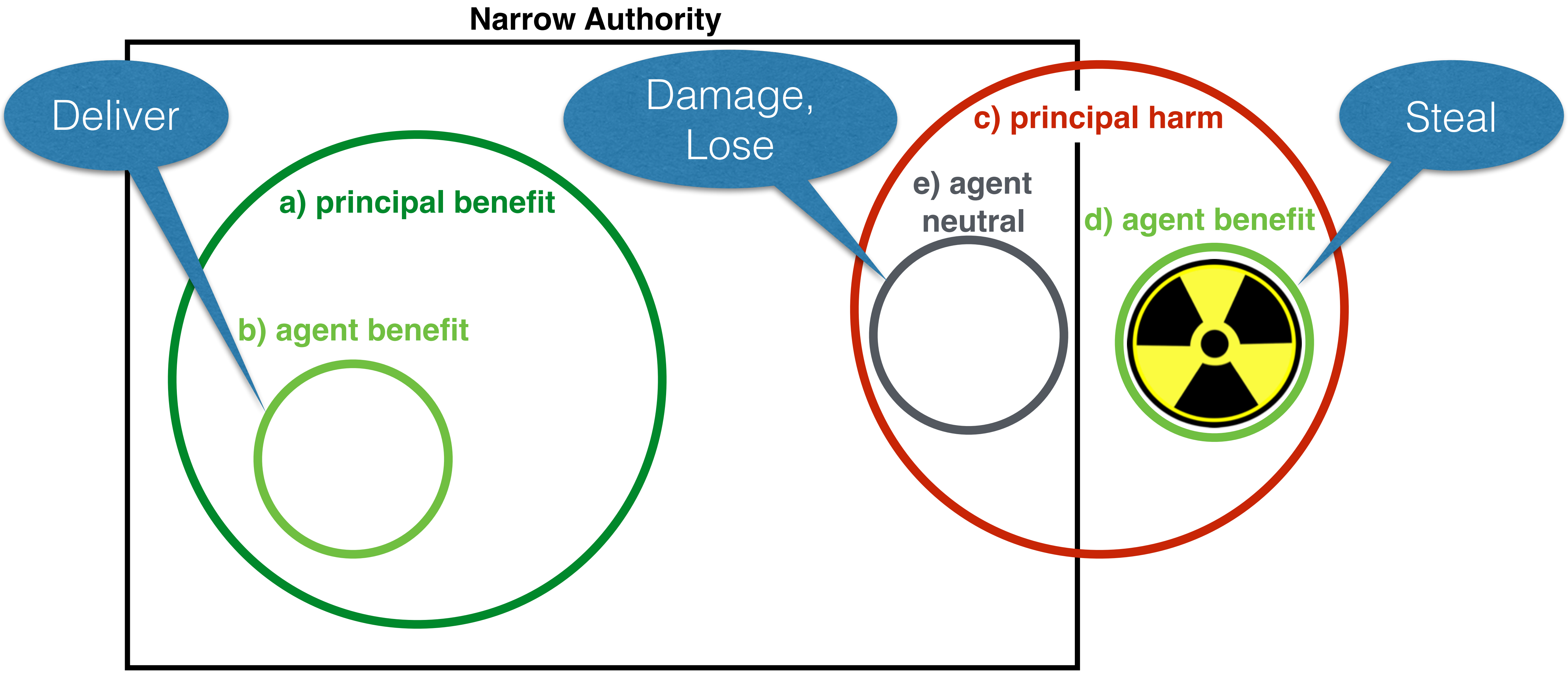
Reward

Broad

Least Authority

Guide

Induce



Narrow Authority

Deliver

Damage, Lose

Steal

a) principal benefit

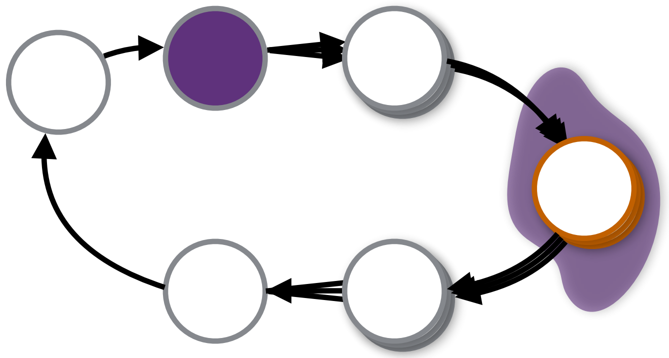
b) agent benefit

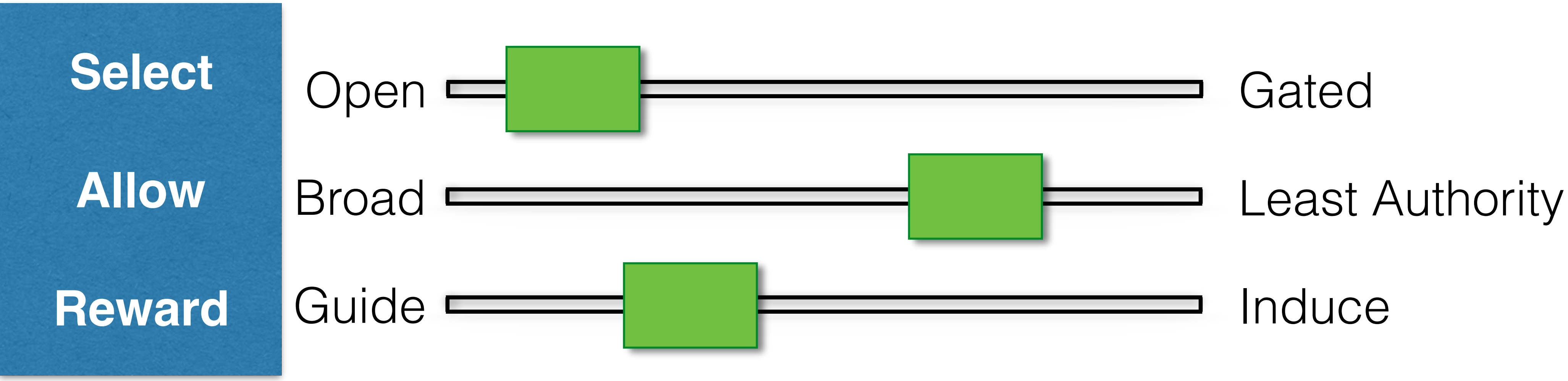
c) principal harm

e) agent neutral

d) agent benefit

Allow + **Reward** + a little **Select**

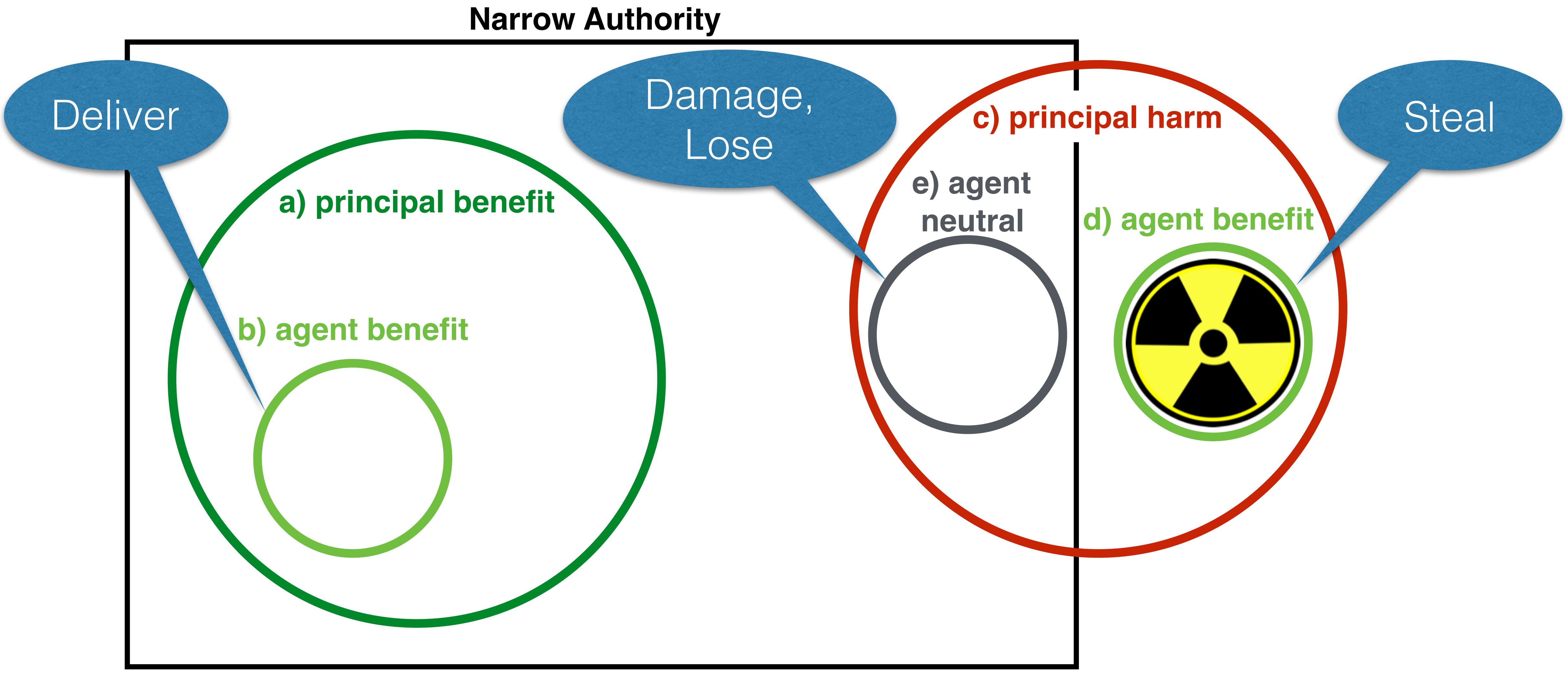




Select Open — Gated

Allow Broad — Least Authority

Reward Guide — Induce



Narrow Authority

Deliver

Damage, Lose

Steal

a) principal benefit

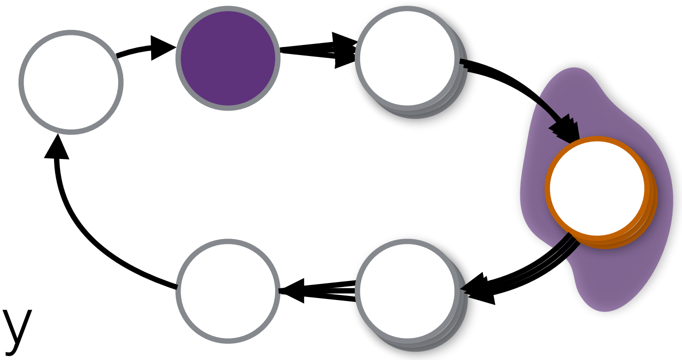
b) agent benefit

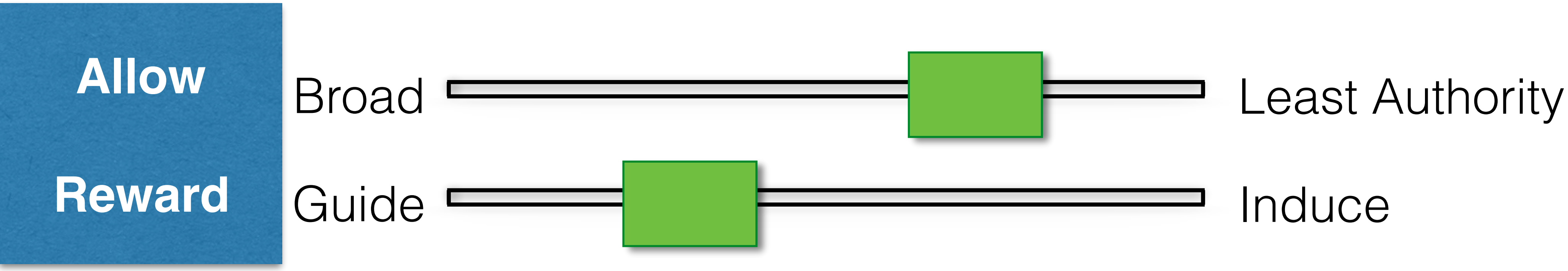
c) principal harm

d) agent benefit

e) agent neutral

Plugin Safety





Allow

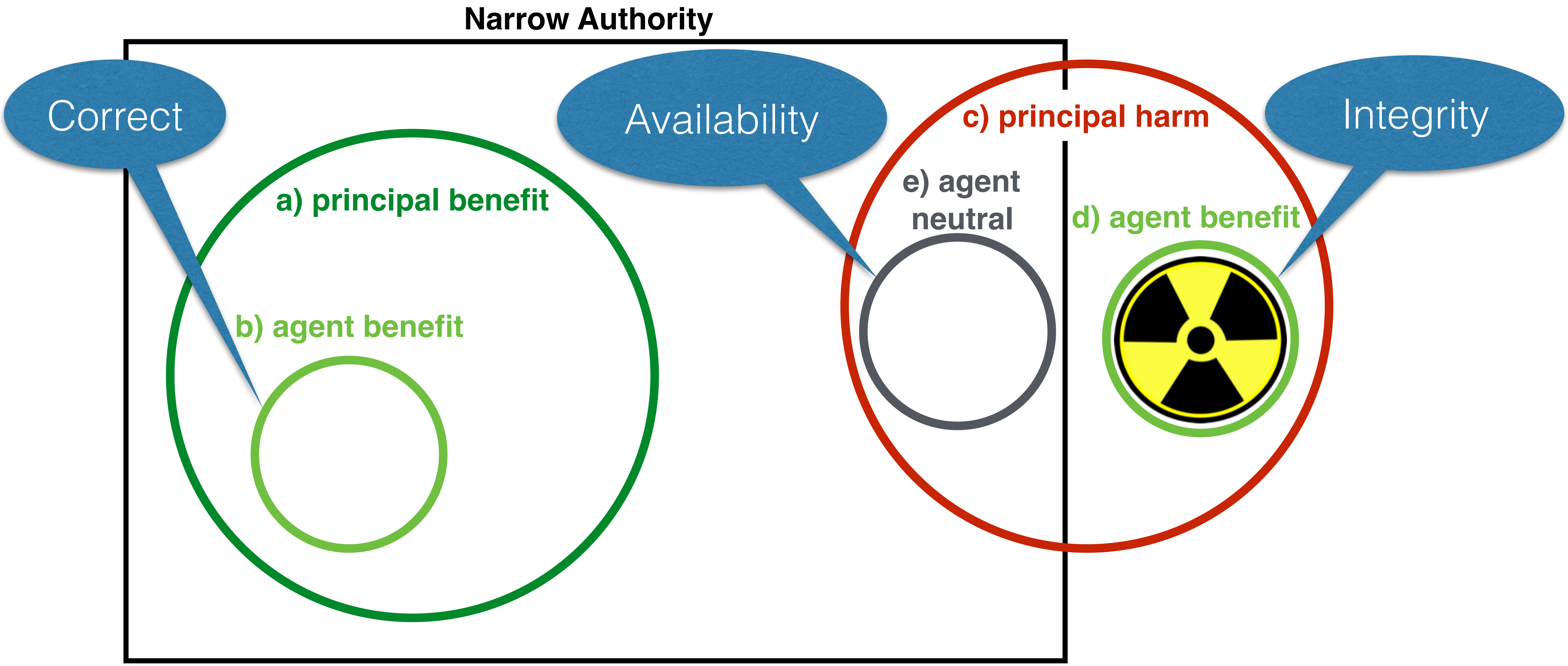
Reward

Broad

Least Authority

Guide

Induce



Narrow Authority

Correct

Availability

Integrity

a) principal benefit

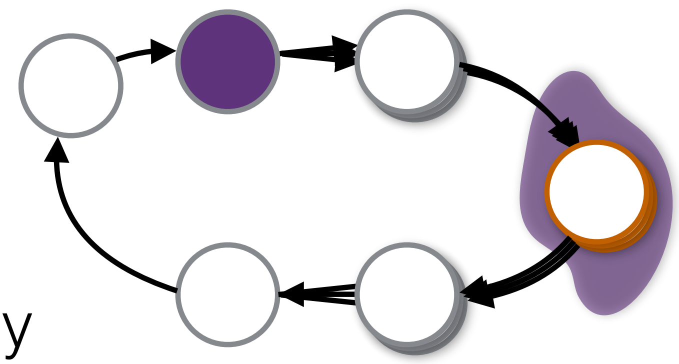
b) agent benefit

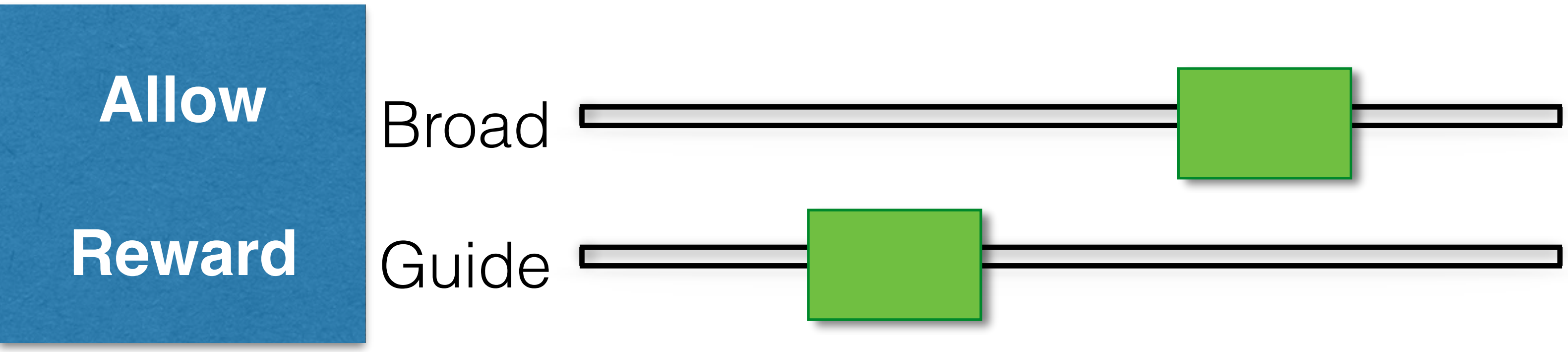
c) principal harm

e) agent neutral

d) agent benefit

Plugin Safety



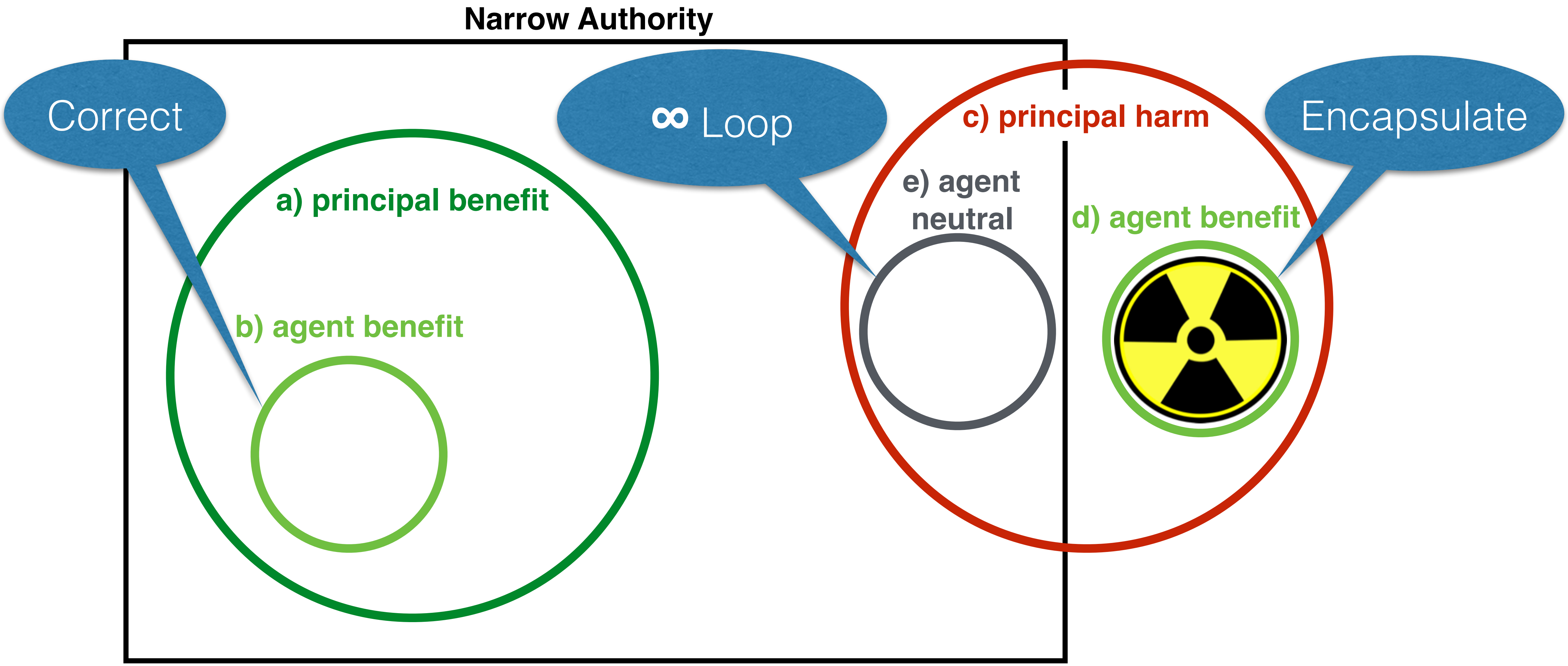


Allow

Broad — Least Authority

Reward

Guide — Induce



Narrow Authority

Correct

a) principal benefit

b) agent benefit

∞ Loop

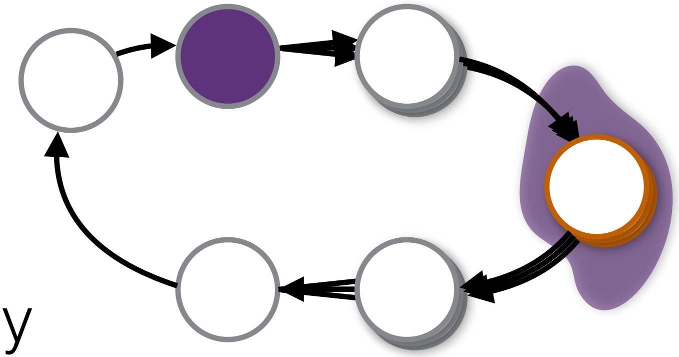
c) principal harm

e) agent neutral

d) agent benefit

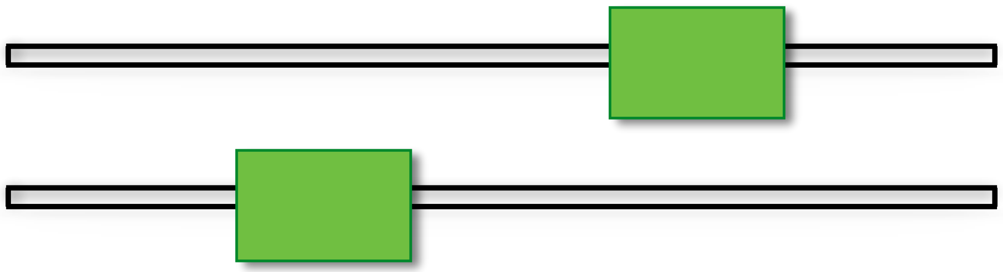
Encapsulate

Allow + Reward



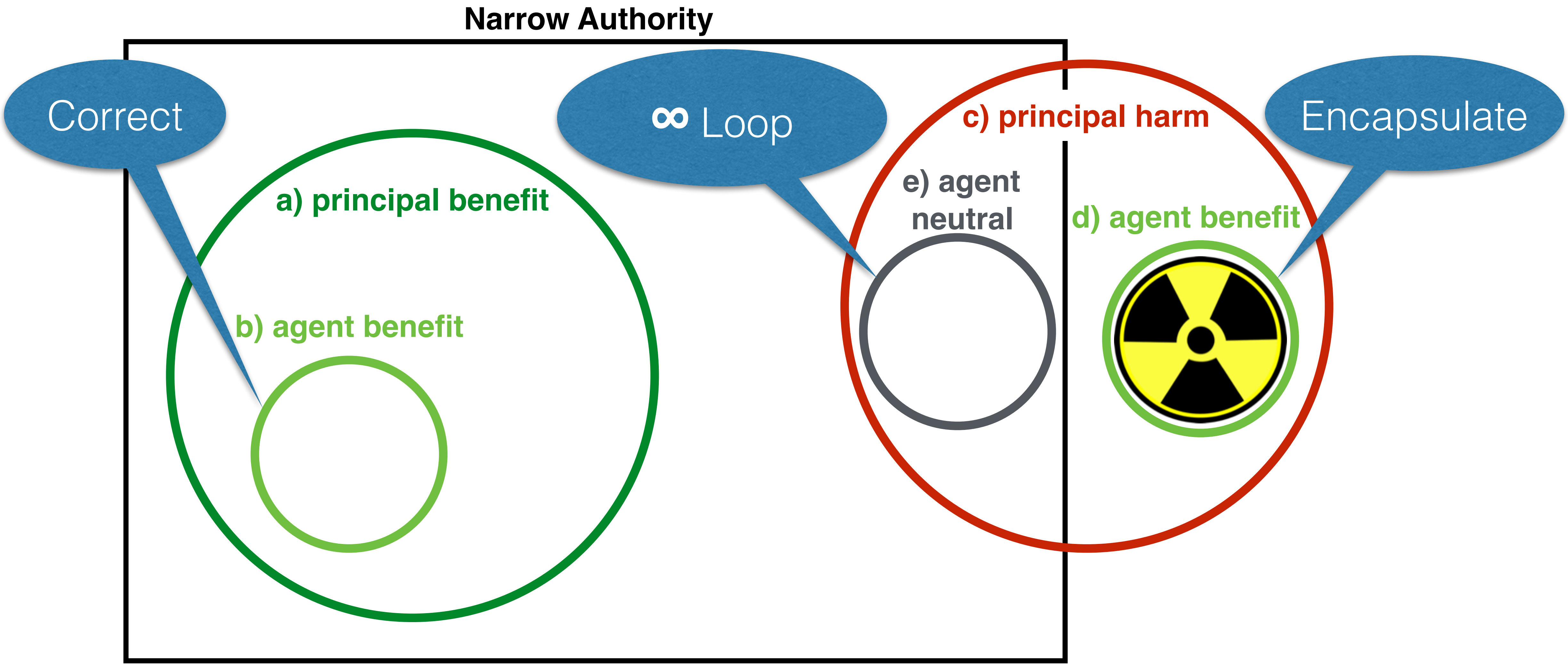
Allow

Reward



Broad — Least Authority

Guide — Induce



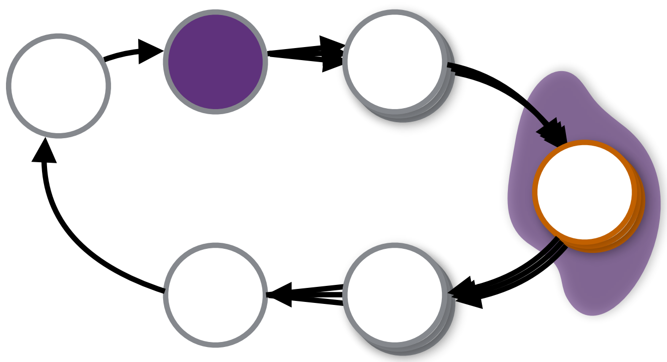
Narrow Authority

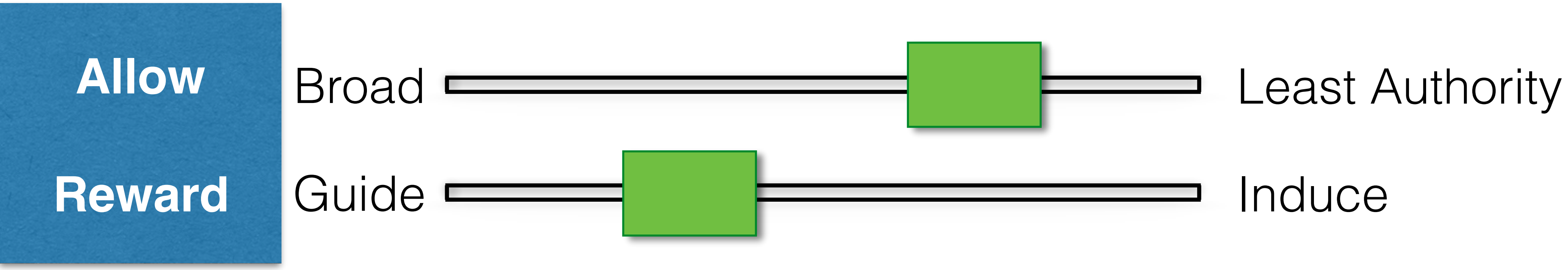
Correct

∞ Loop

Encapsulate

Allow + Reward





Allow

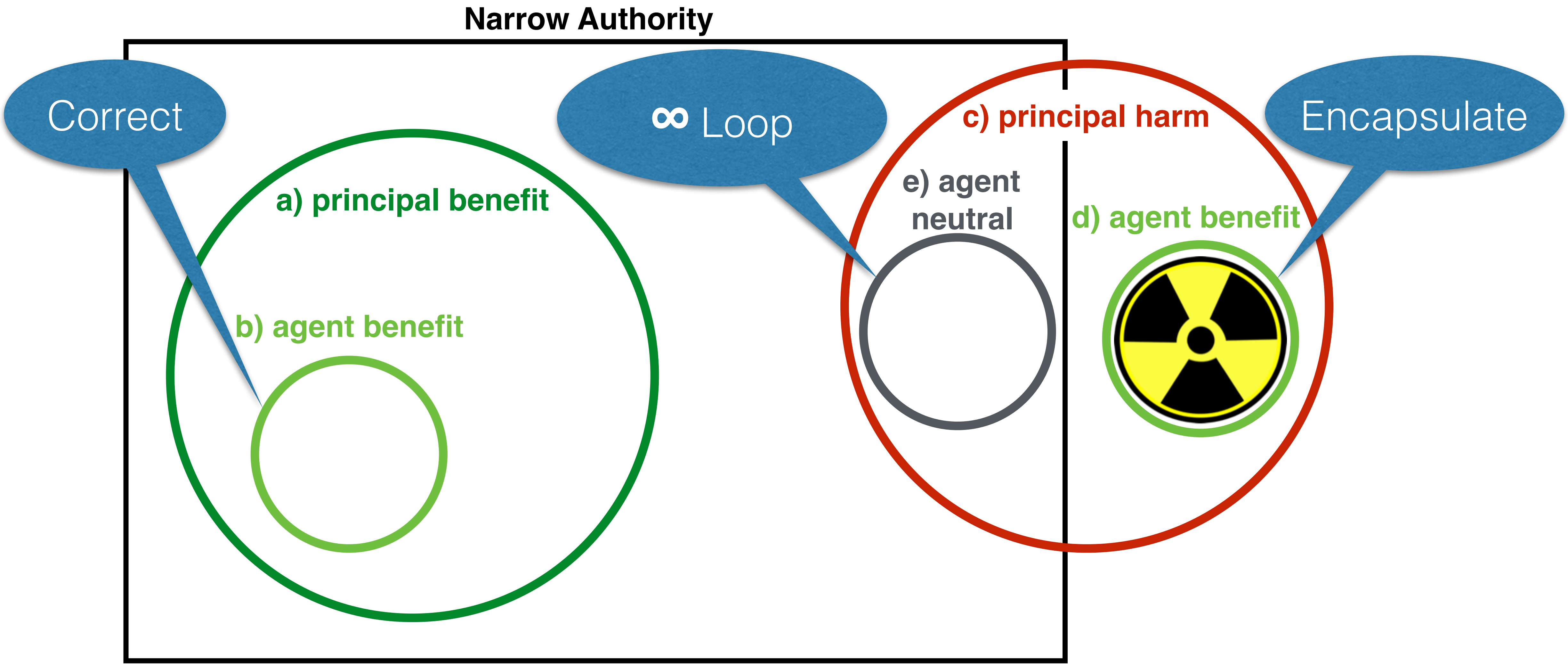
Broad

Least Authority

Reward

Guide

Induce



Narrow Authority

Correct

∞ Loop

Encapsulate

a) principal benefit

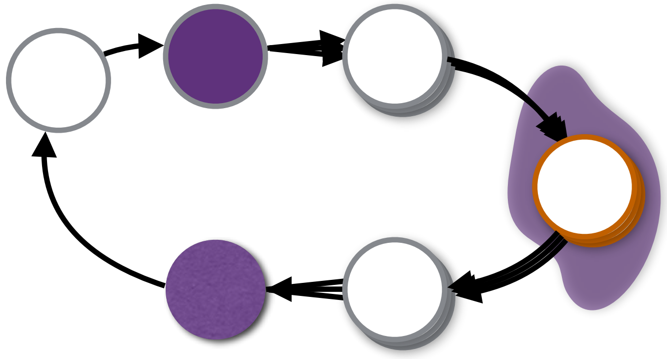
b) agent benefit

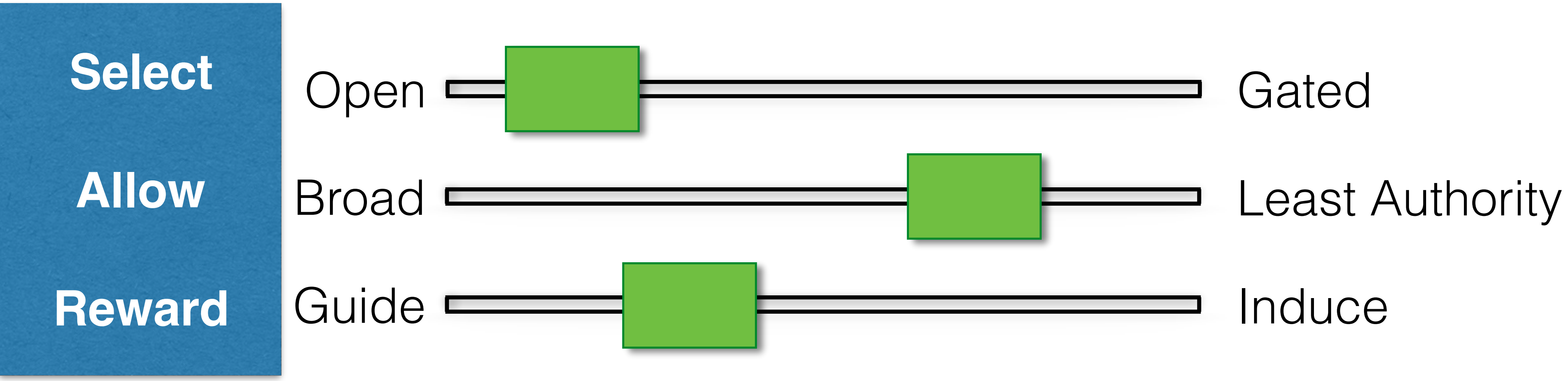
c) principal harm

e) agent neutral

d) agent benefit

Allow + **Reward** + a little **Select**





Select

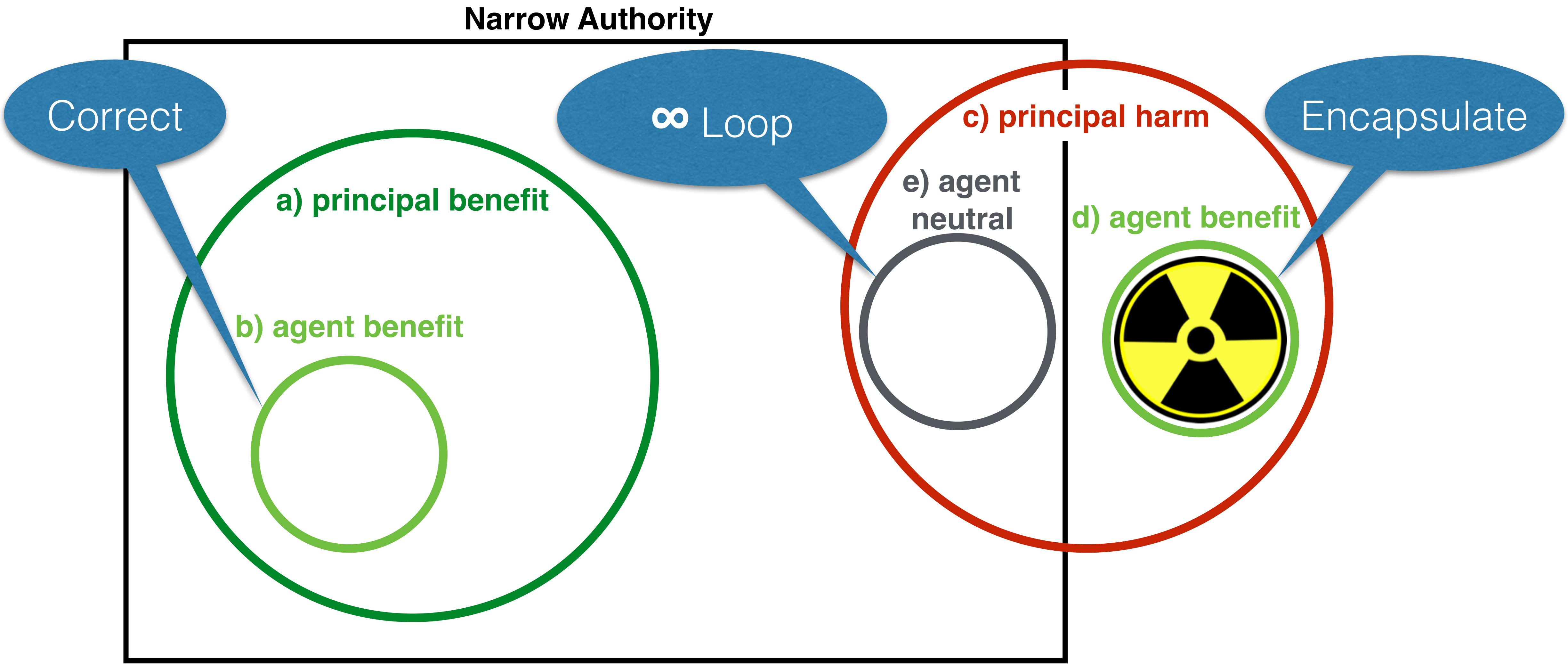
Allow

Reward

Open — Gated

Broad — Least Authority

Guide — Induce



Narrow Authority

Correct

∞ Loop

Encapsulate

a) principal benefit

b) agent benefit

c) principal harm

d) agent benefit

e) agent neutral

Spectrum of Allowing Effects

Isolating effects in Space

Memory-unsafe, imperative

Spectrum of Allowing Effects

Isolating effects in Space

Memory-unsafe, imperative

Purely Functional

Spectrum of Allowing Effects

Isolating effects in Space

Memory-unsafe, imperative

Memory-safe, imperative

Purely Functional

Spectrum of Allowing Effects

Isolating effects in Space

Memory-unsafe, imperative

Memory-safe, imperative

OCap

Purely Functional

Spectrum of Allowing Effects

Isolating effects in Space

Memory-unsafe, imperative

Memory-safe, imperative

OCap

Purely Functional

Isolating effects in Time

Sequential programming

Spectrum of Allowing Effects

Isolating effects in Space

Memory-unsafe, imperative

Memory-safe, imperative

OCap

Purely Functional

Isolating effects in Time

Pre-emptive multithreading

Sequential programming

Spectrum of Allowing Effects

Isolating effects in Space

Memory-unsafe, imperative

Memory-safe, imperative

OCap

Purely Functional

Isolating effects in Time

Pre-emptive multithreading

Cooperative multithreading

Sequential programming

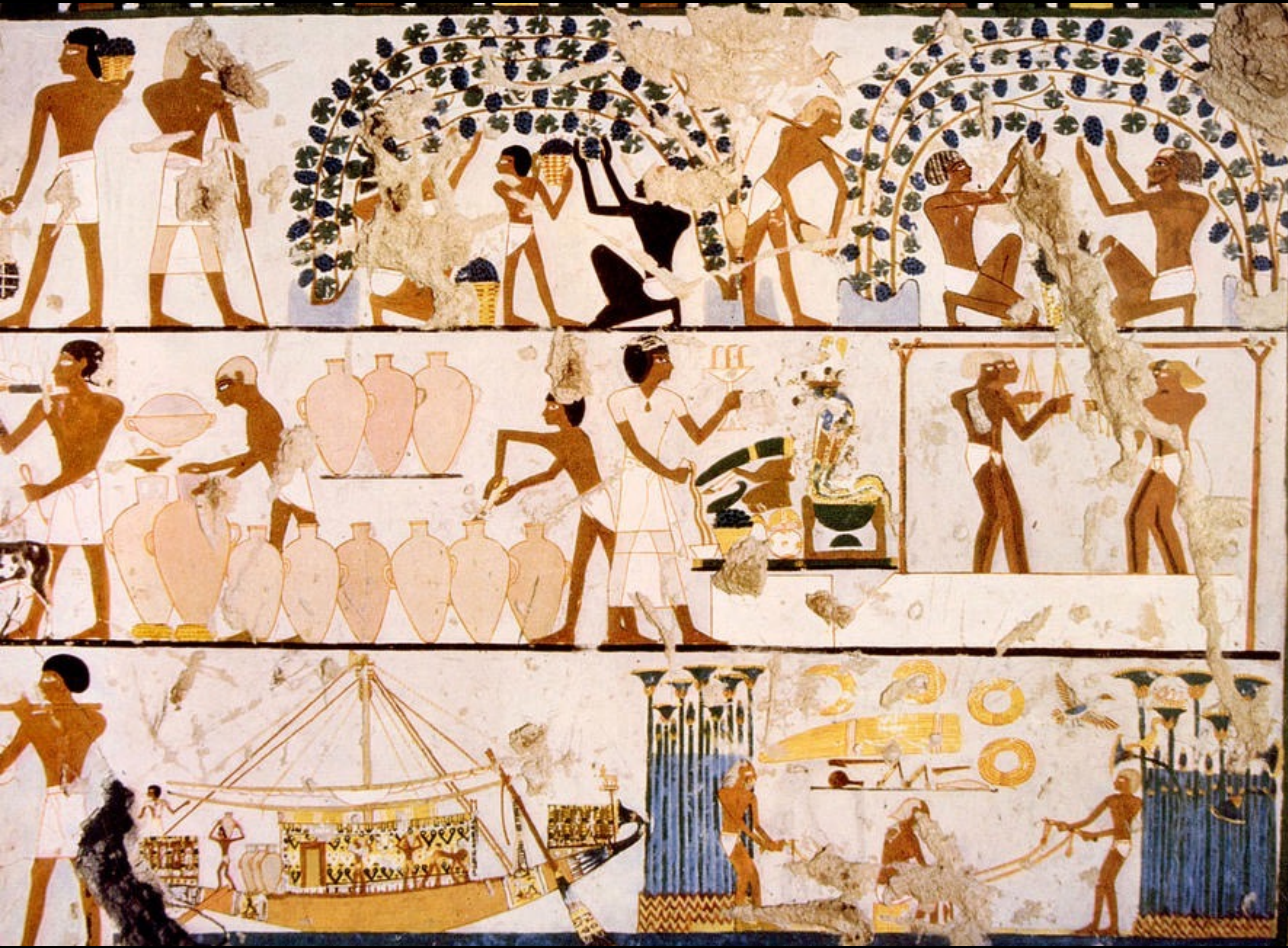
Spectrum of Allowing Effects

Isolating effects in Space

- Memory-unsafe, imperative
- Memory-safe, imperative
- OCap
- Purely Functional

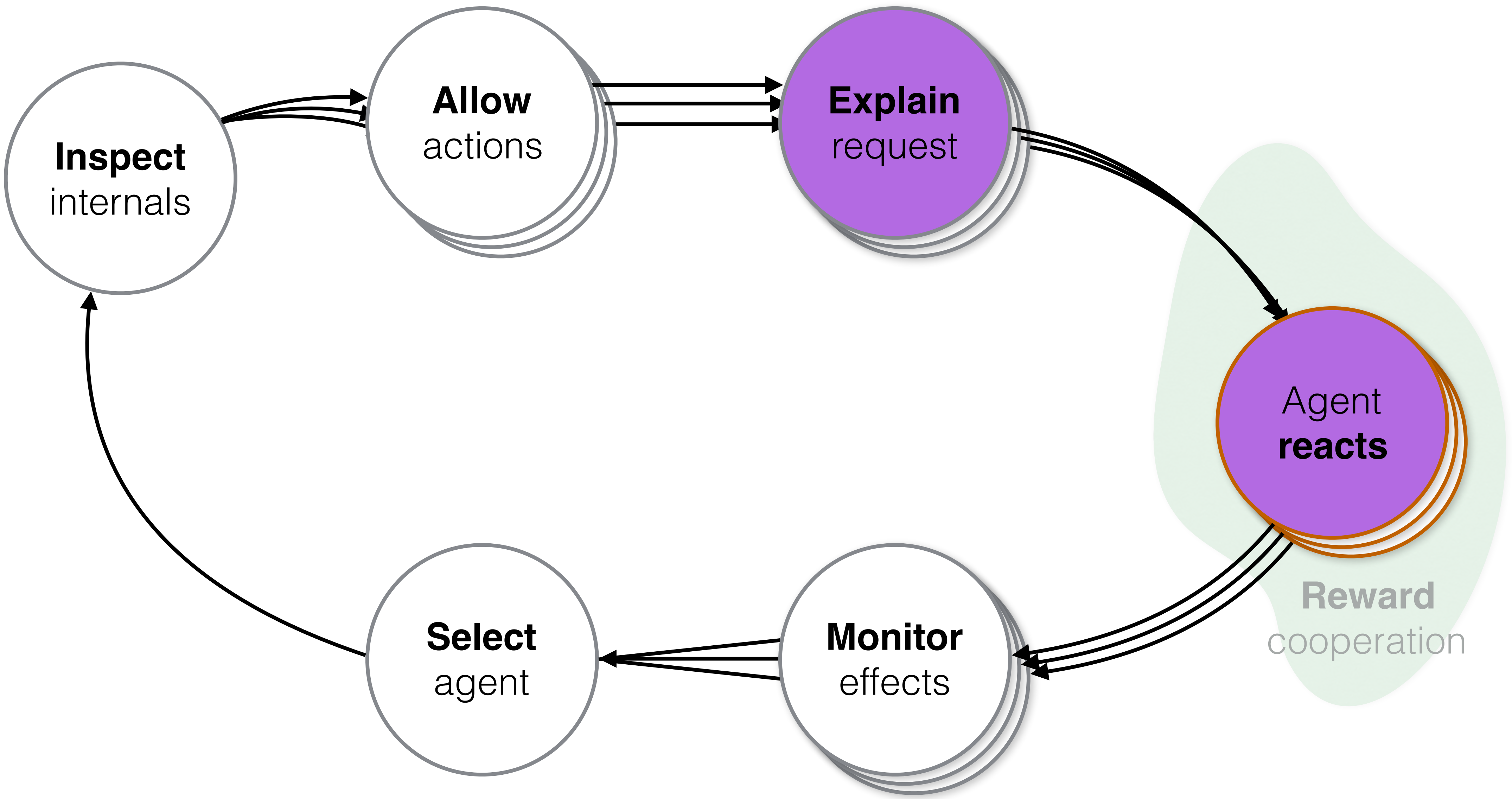
Isolating effects in Time

- Pre-emptive multithreading
- Cooperative multithreading
- **Communicating Event Loops**
- Sequential programming

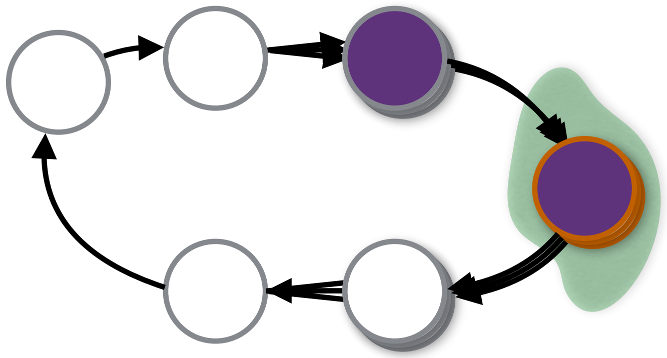


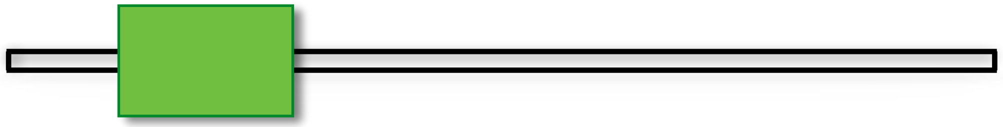
Internal Software Development

API Design



Explain request
Shared Understandings



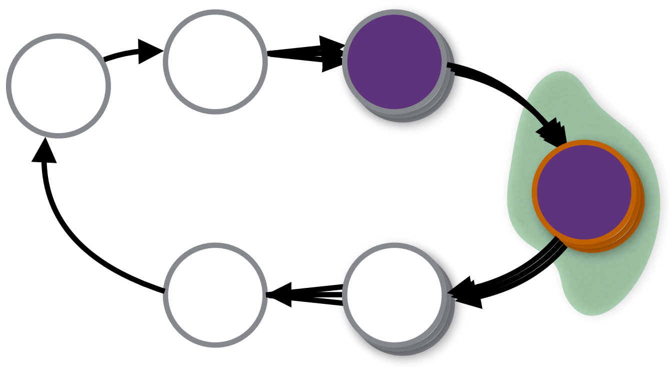


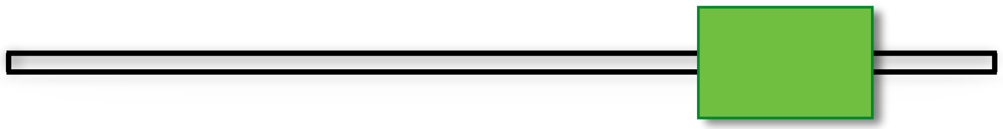
Informal

Specified

Principal *why*	Parsing	Gift for Dad
Interface *what*	parens[i++] = tok	“Give this to my Dad?”
Agent *how*	Array+index	Plane

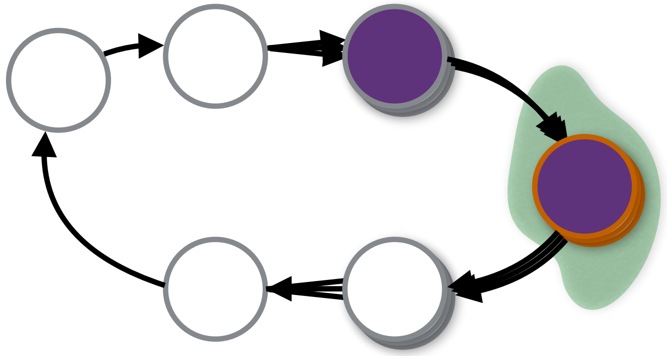
Explain request
Shared Understandings

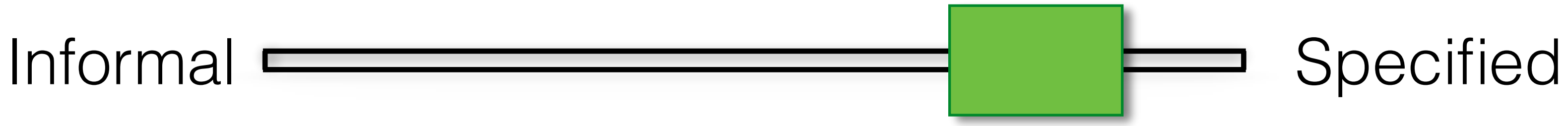


Informal  Specified

Principal *why*	Parsing	Gift for Dad
Interface *what*	*Stack*	*Package delivery*
Agent *how*	Array+index	Plane

Explain request
Abstraction boundaries





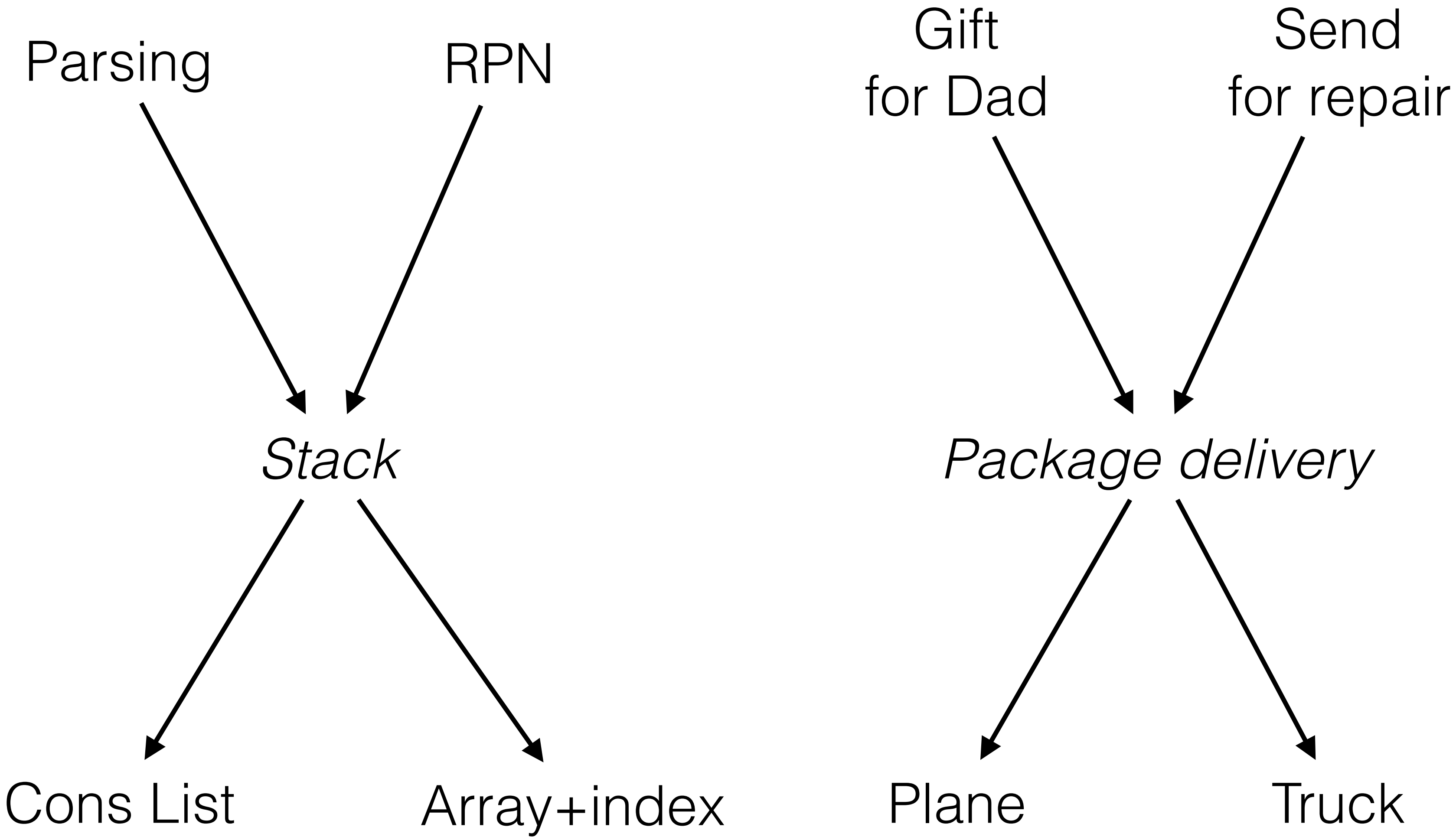
Informal

Specified

Principal *why*
Multiple purposes

Interface *what*

Agent *how*
Multiple means



Parsing

RPN

Gift
for Dad

Send
for repair

Stack

Package delivery

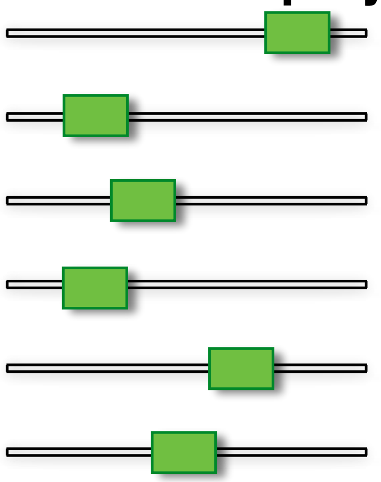
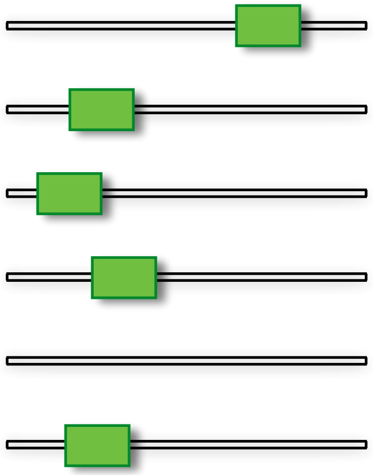
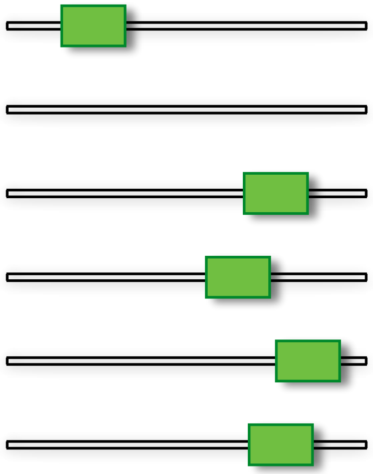
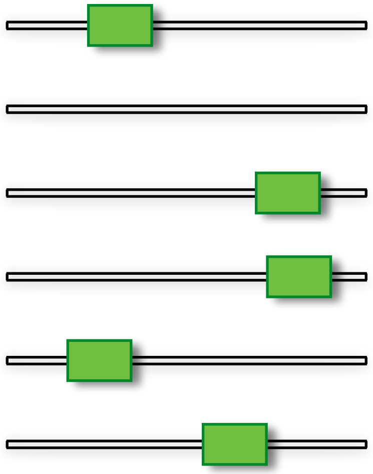
Cons List

Array+index

Plane

Truck

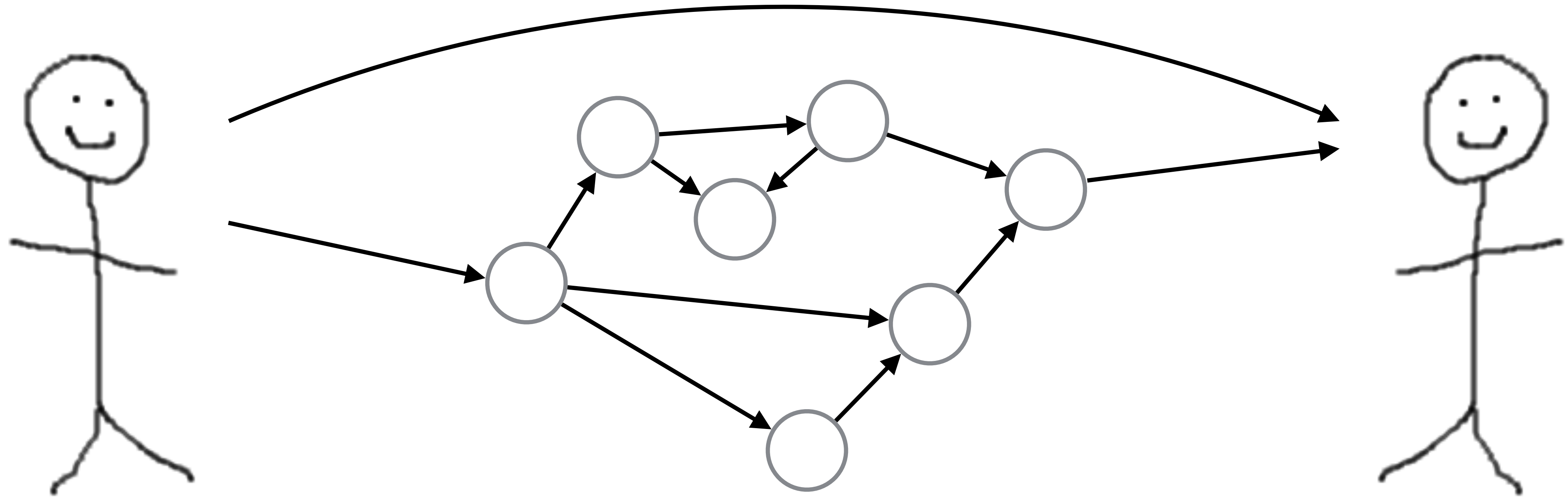
Structural Similarities

	Human to Human	Object to Object
Gated Hierarchy Trusting Informal Concrete	Organizational Employment 	Internal Software Development 
Open Network Defensive Specified Abstract	Package Delivery Business 	Safe Plugin Boundary 

Networks of Specialized Knowledge

Networks of Specialized Knowledge

Shifting mixtures of humans and software.



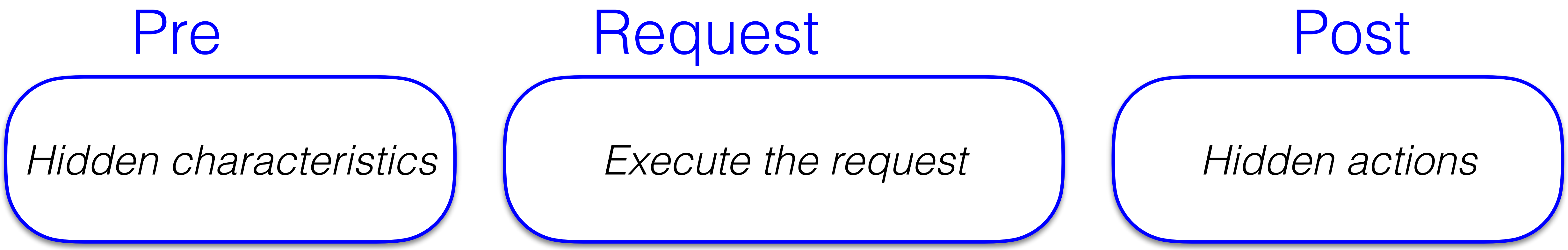
Networks of Specialized Knowledge

Shifting mixtures of humans and software.

Division of knowledge hazards

Econ: intentional misbehavior

CS: accidental misbehavior



Pre

Hidden characteristics

Request

Execute the request

Post

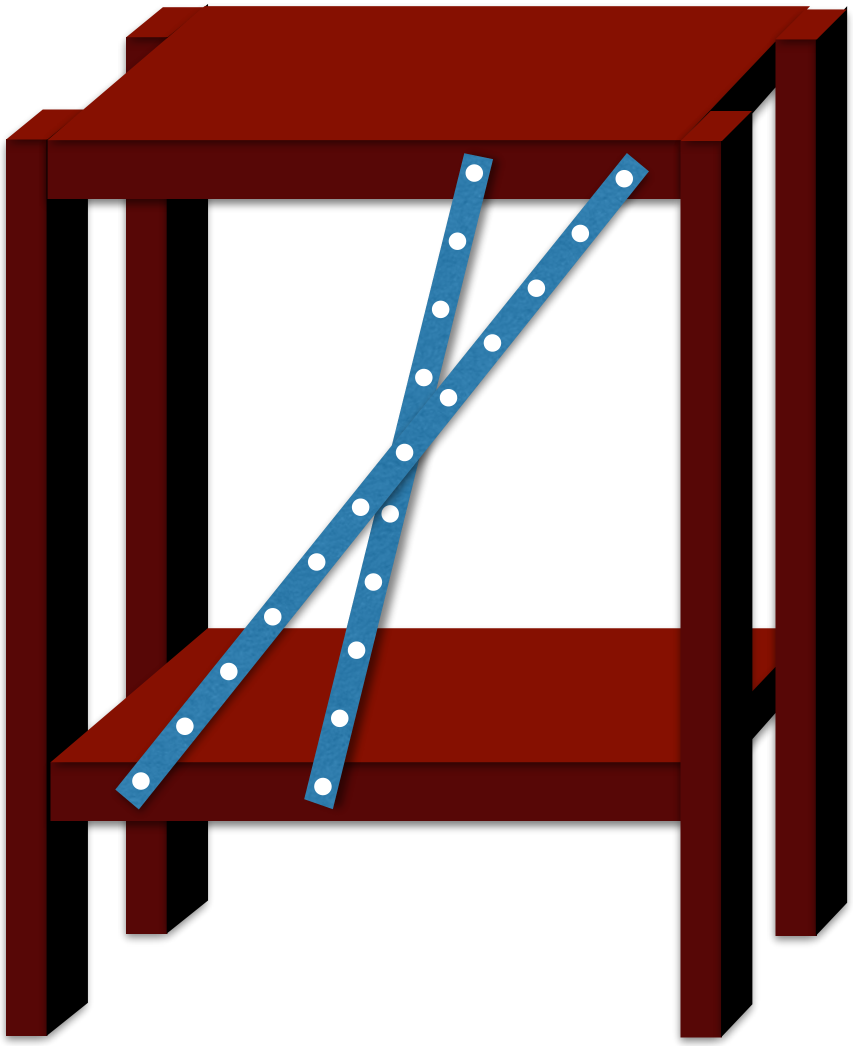
Hidden actions

Networks of Specialized Knowledge

Shifting mixtures of humans and software.

Division of knowledge hazards
  Econon: intentional misbehavior
  CS: accidental misbehavior

Compose compromises
  Study and support

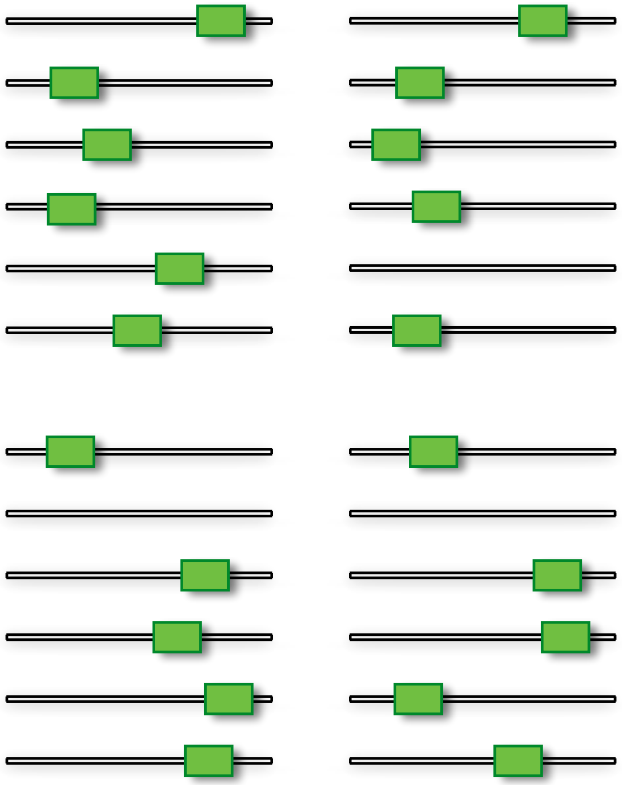


Networks of Specialized Knowledge

Shifting mixtures of humans and software.

Division of knowledge hazards
    Econ: intentional misbehavior
    CS: accidental misbehavior

Compose compromises
    Study and support

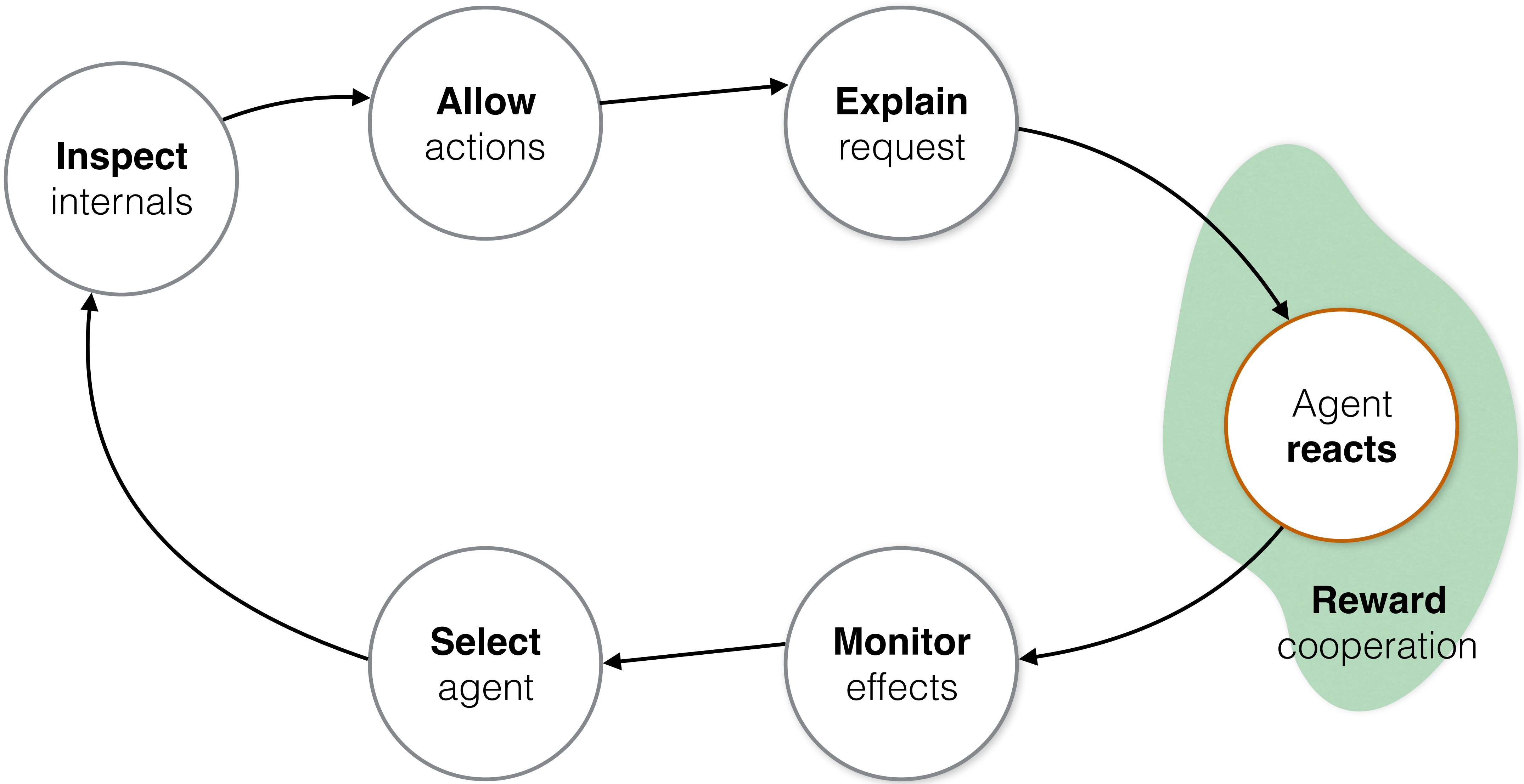


Emergent properties when things go right
and when things go wrong.

Questions?

Danger
Oversimplifications
Ahead

The Principal-Agent Loop



Inspect internals

Allow actions

Explain request

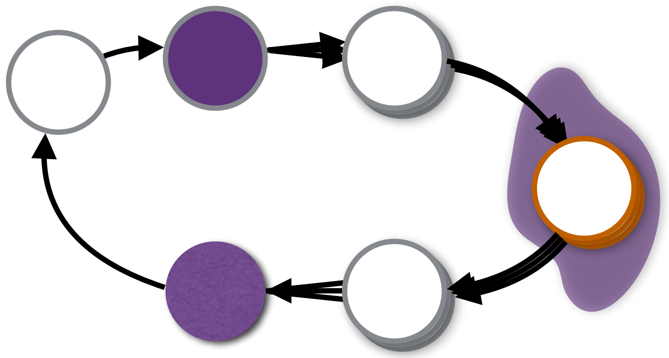
Agent **reacts**

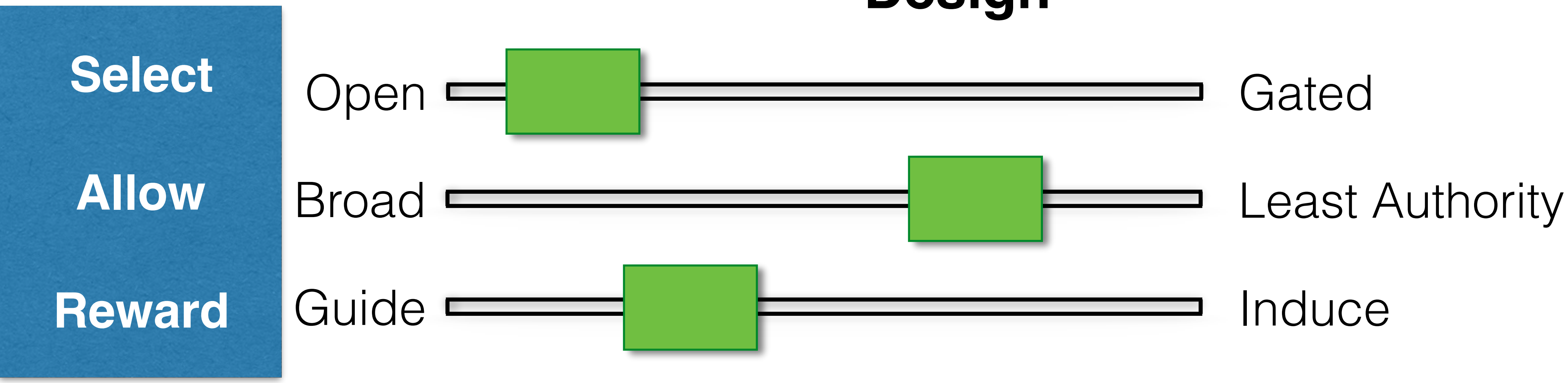
Reward cooperation

Monitor effects

Select agent

Secure Language Design





Select

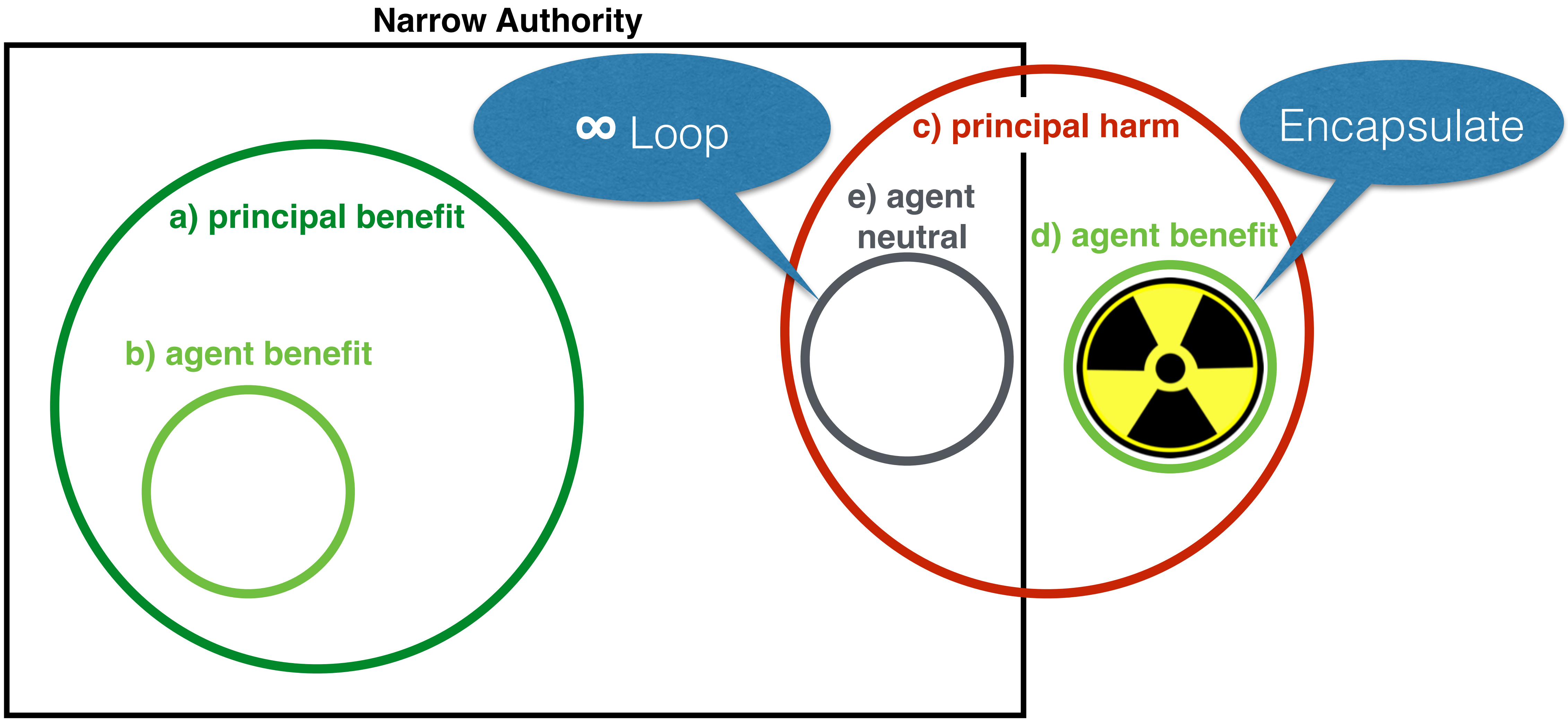
Allow

Reward

Open — Gated

Broad — Least Authority

Guide — Induce



Narrow Authority

a) principal benefit

b) agent benefit

∞ Loop

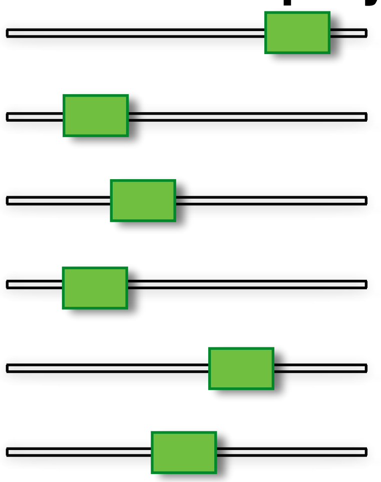
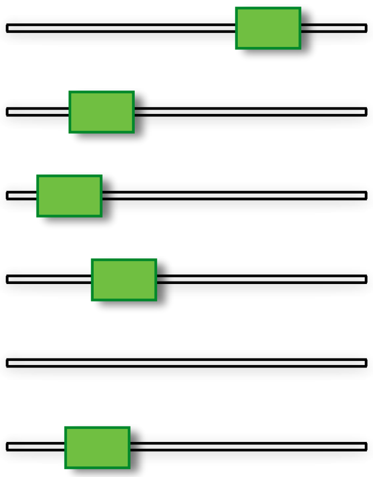
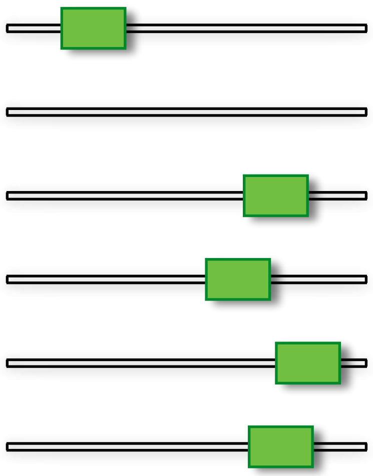
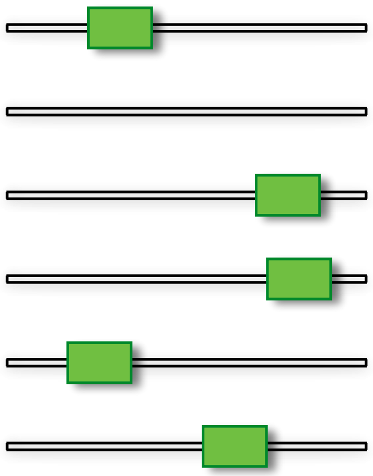
c) principal harm

e) agent neutral

d) agent benefit

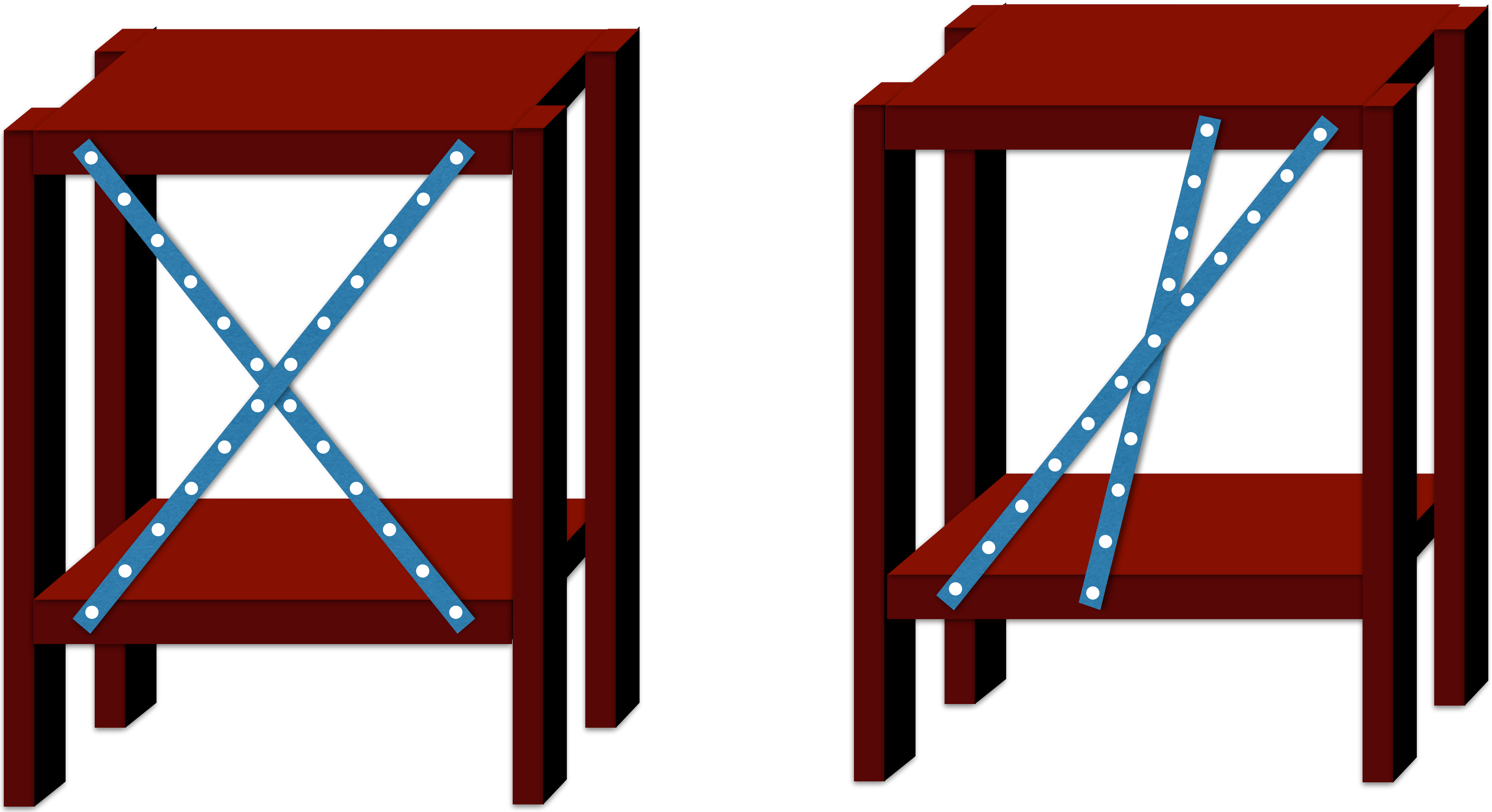
Encapsulate

Structural Similarities

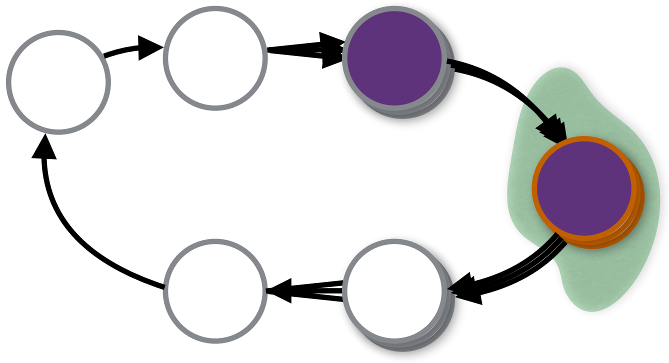
	Human to Human	Object to Object
Gated Hierarchy Trusting Informal Concrete	Organizational Employment 	Internal Software Development 
Open Network Defensive Specified Abstract	Package Delivery Business 	Safe Plugin Boundary 

Cross Bracing

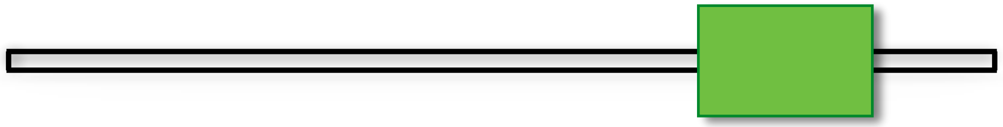
Co-design. Joint application.



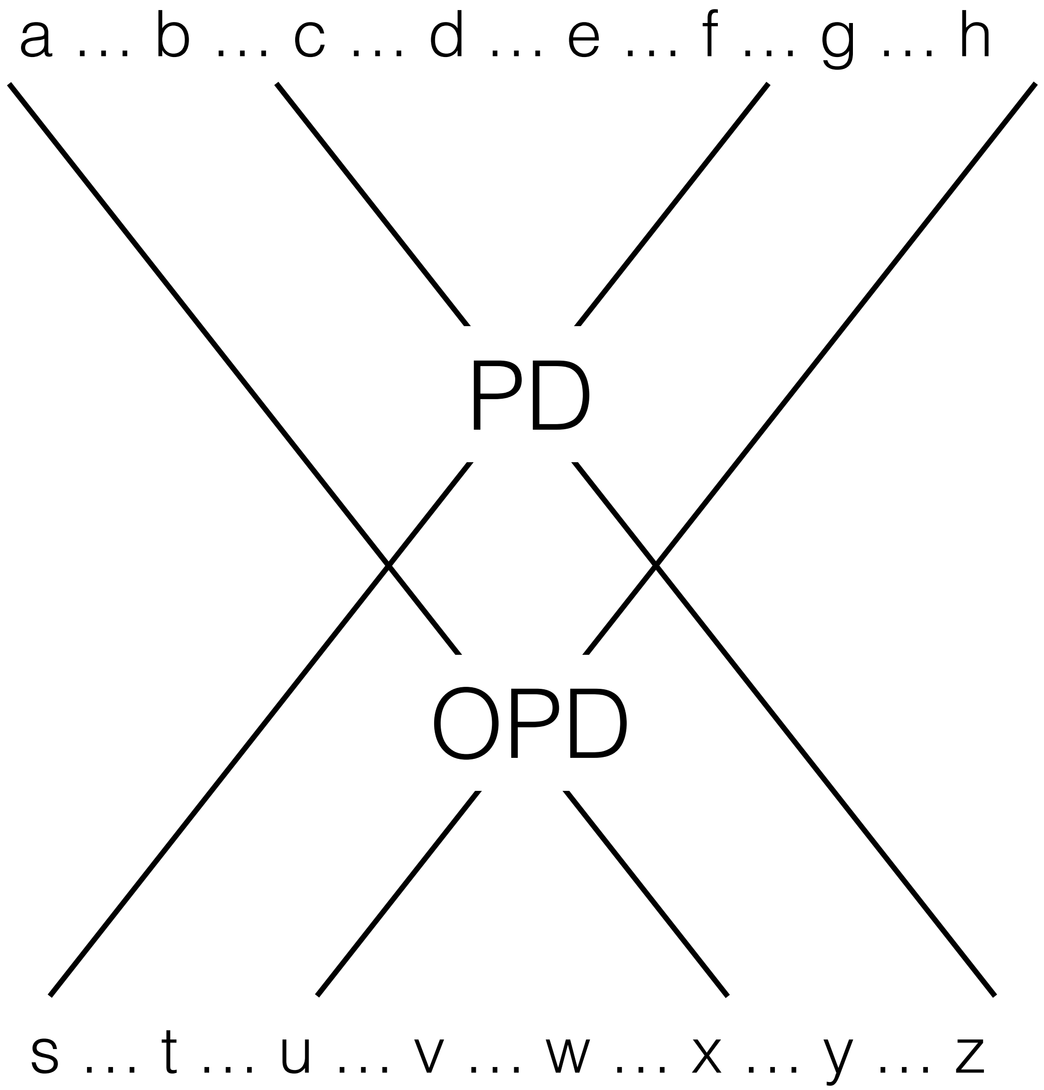
Explain request
Which is “Wider”?



Informal



Specified



Principal *why*
Multiple purposes

a … b … c … d … e … f … g … h

Supertype

PD

Package Delivery

Subtype

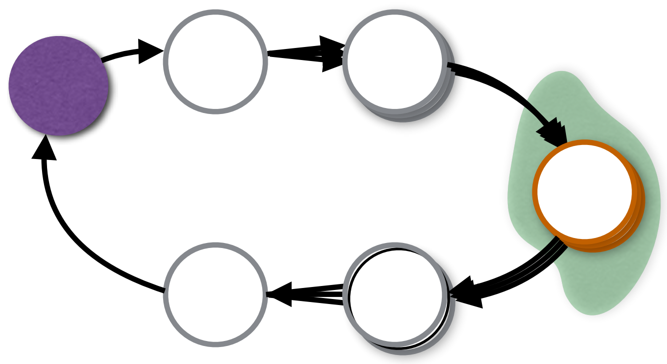
OPD

Overnight Package Delivery

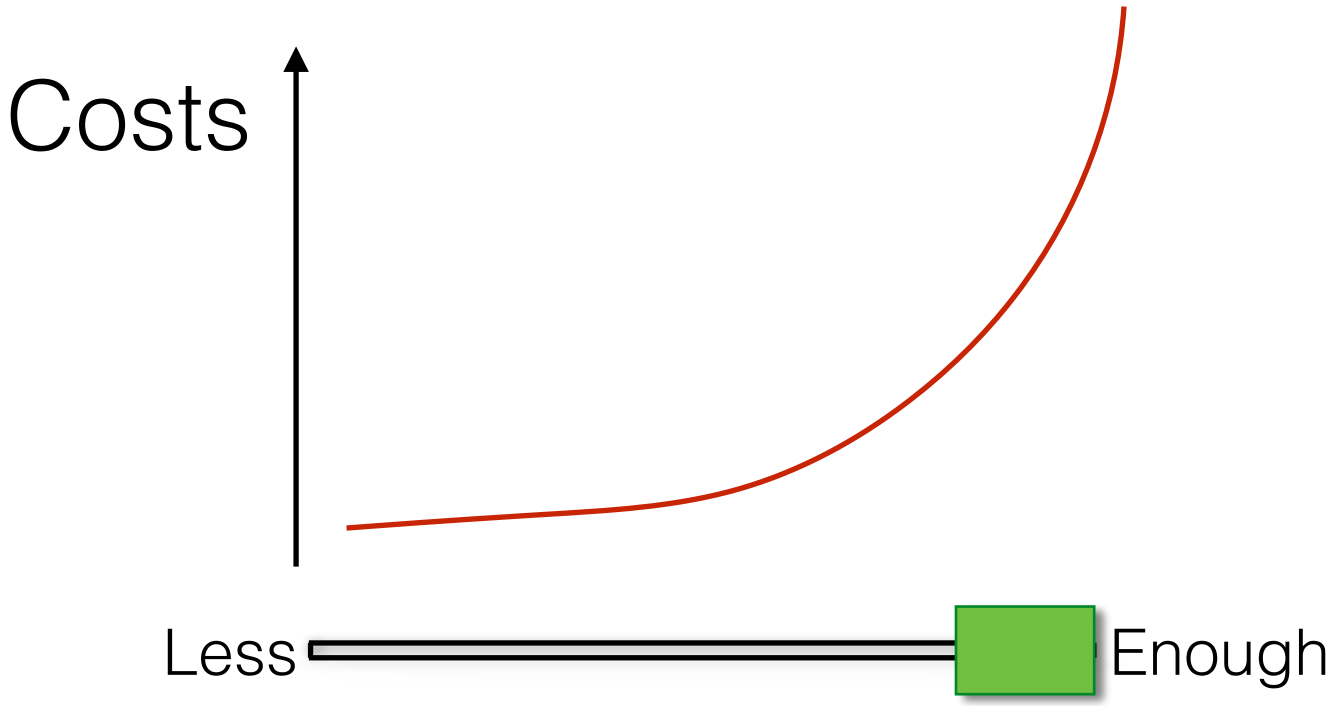
Agent *how*
Multiple means

s … t … u … v … w … x … y … z

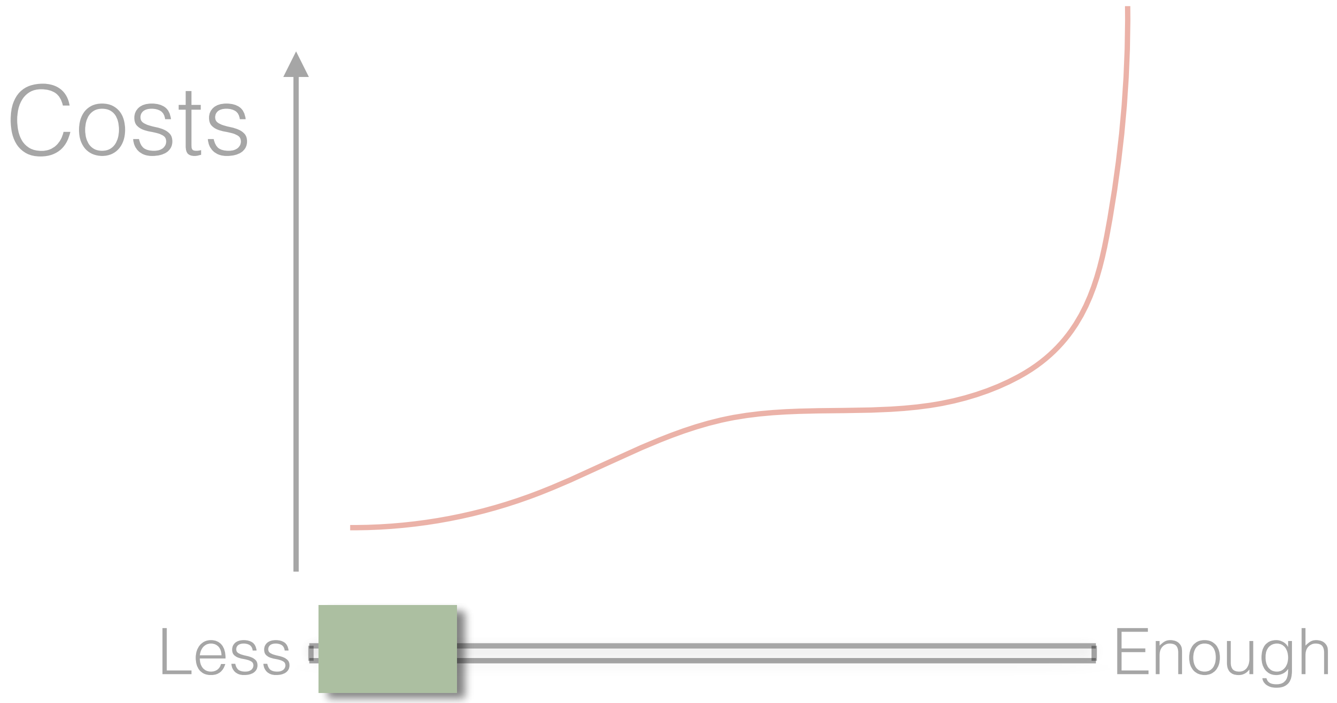
Academia *vs* Industry



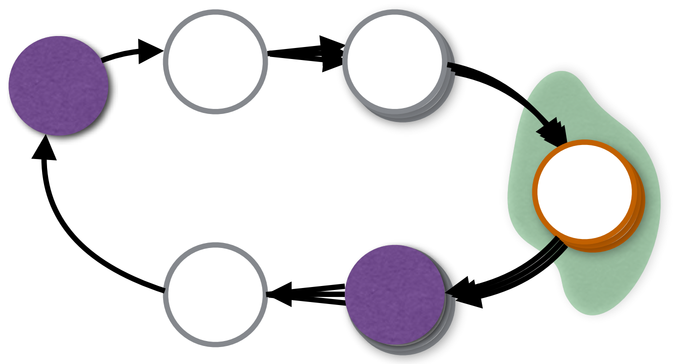
Inspect



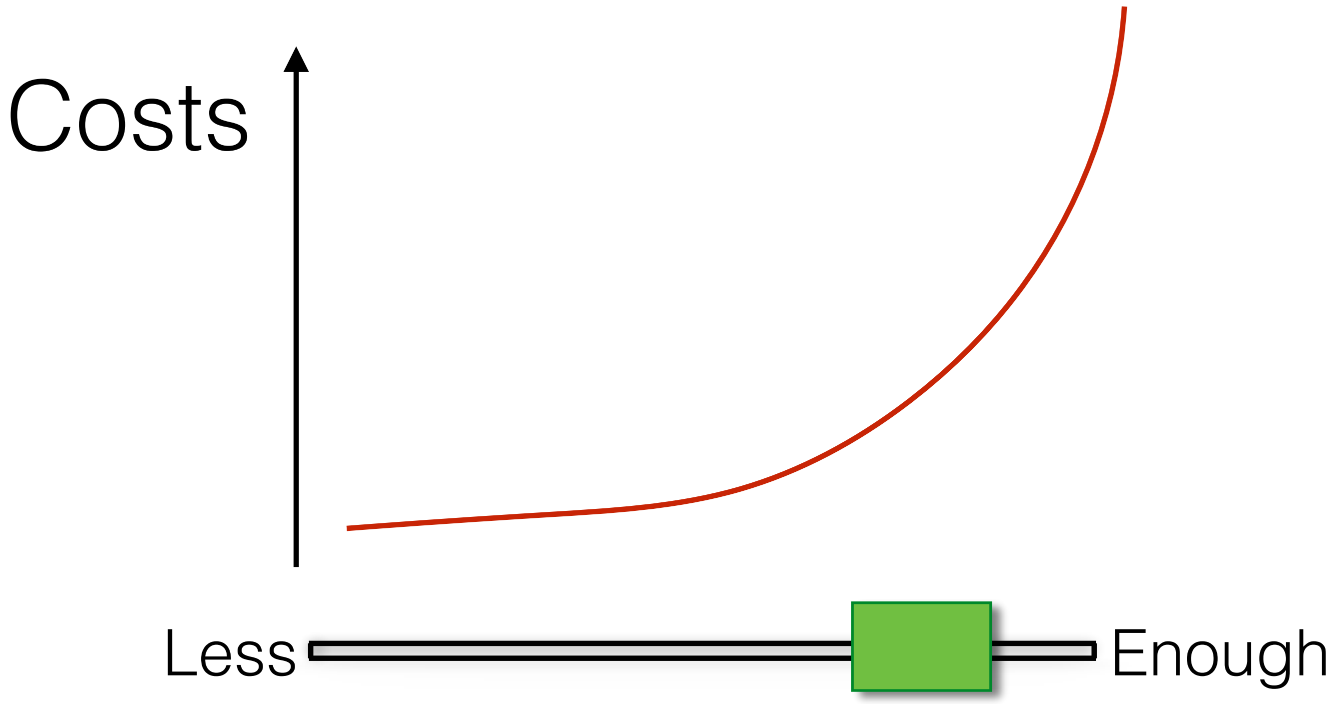
Monitor



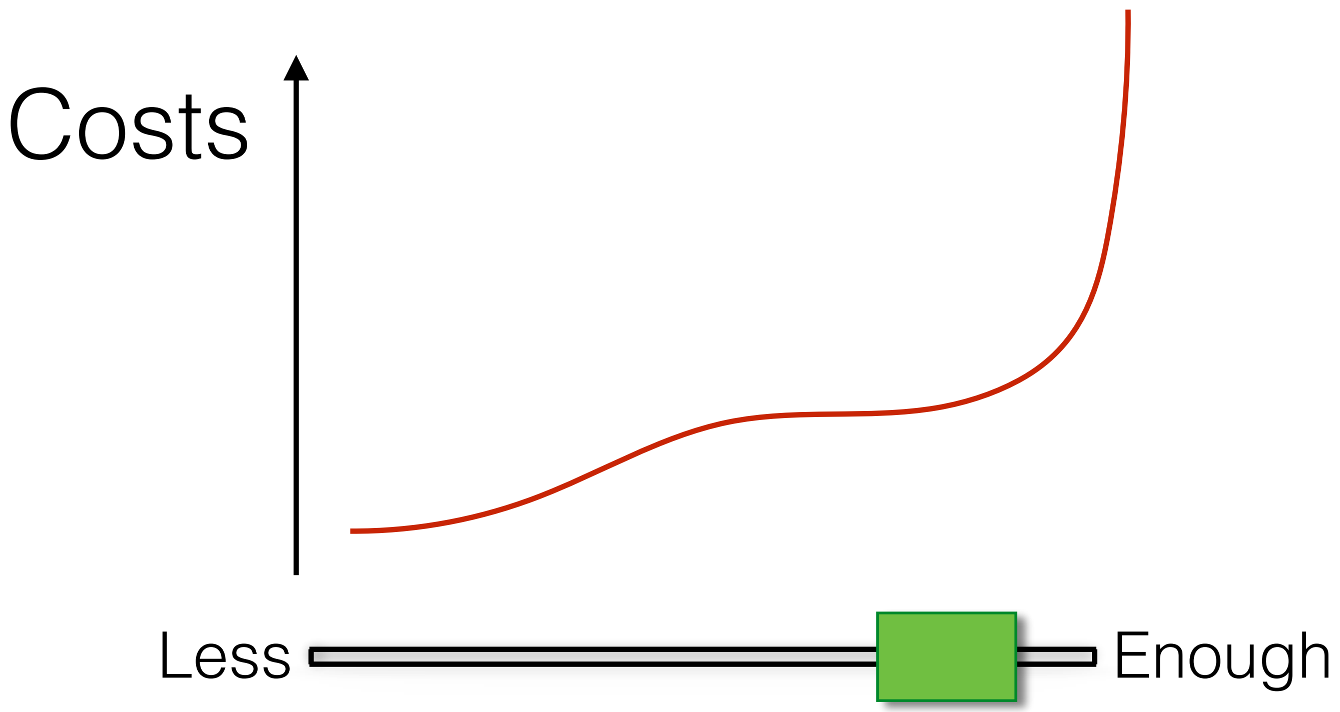
Academia *vs* Industry



Inspect



Monitor



Cooperation with less Vulnerability

My journey

	+Cooperation	-Vulnerability
Xanadu Hypertext		
Agoric Open Systems		
Object-capabilities		
JavaScript standards Frozen realms		

Cooperation with less Vulnerability

My journey

	+Cooperation	**-Vulnerability**
Xanadu Hypertext	**Bi-directional links**	**Fine-grained skepticism**
Agoric Open Systems		
Object-capabilities		
JavaScript standards Frozen realms		

X

Cooperation with less Vulnerability

My journey

	+Cooperation	**-Vulnerability**
Xanadu Hypertext	Bi-directional links	Fine-grained skepticism
Agoric Open Systems	Prices guide tradeoffs	**Encapsulation as property rights**
Object-capabilities		
JavaScript standards Frozen realms		

Cooperation with less Vulnerability

My journey

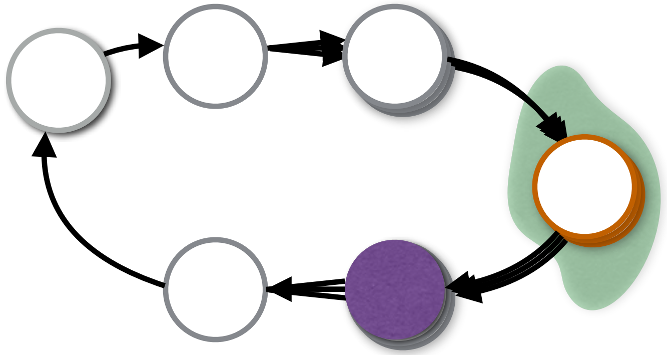
	+Cooperation	**-Vulnerability**
Xanadu Hypertext	Bi-directional links	Fine-grained skepticism
Agoric Open Systems	Prices guide tradeoffs	Encapsulation as property rights
Object-capabilities	Authority-driven design	**Nothing but objects**
JavaScript standards Frozen realms		

Cooperation with less Vulnerability

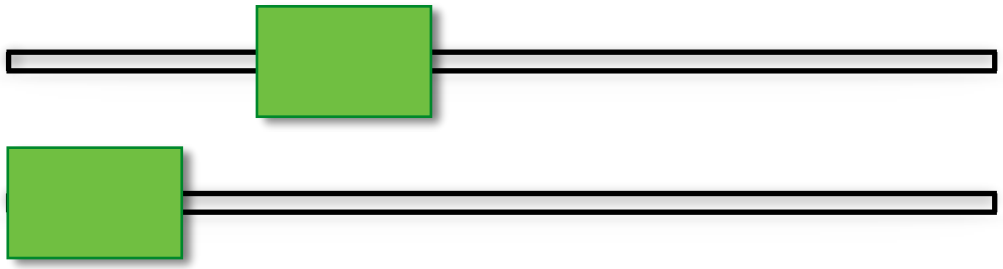
My journey

	+Cooperation	**-Vulnerability**
Xanadu Hypertext	Bi-directional links	Fine-grained skepticism
Agoric Open Systems	Prices guide tradeoffs	Encapsulation as property rights
Object-capabilities	Authority-driven design	Nothing but objects
JavaScript standards **Frozen realms**	**Solid abstraction mechanisms**	**Defend invariants**

Inspect + Monitor



Code Review



Verify

Prevent

Repair Damage

"… tested thoroughly … written specifically to evade testing.

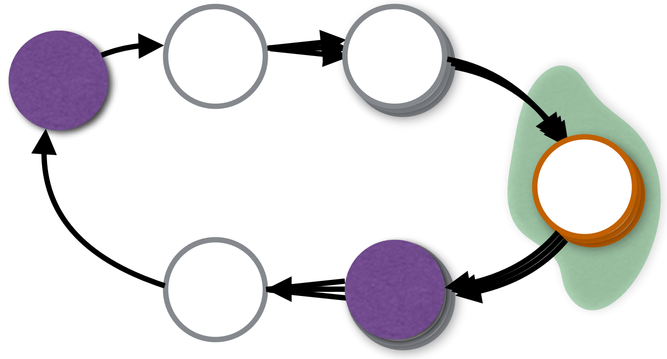
But such evasion is likely to require some suspicious-looking code, …"

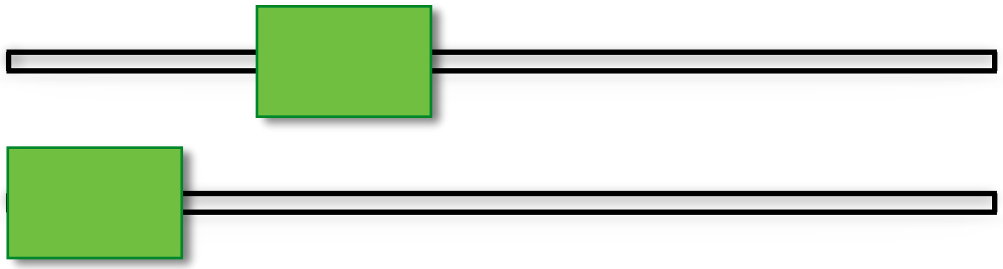
Building Reliable Voting Machine Software
Ka-Ping Yee 2007 dissertation

```
for (i = 0; i < limit; i++)
  if (j === 72374928)
```

Passes all tests.

Inspect + Monitor





Code Review — Verify

Prevent — Repair Damage

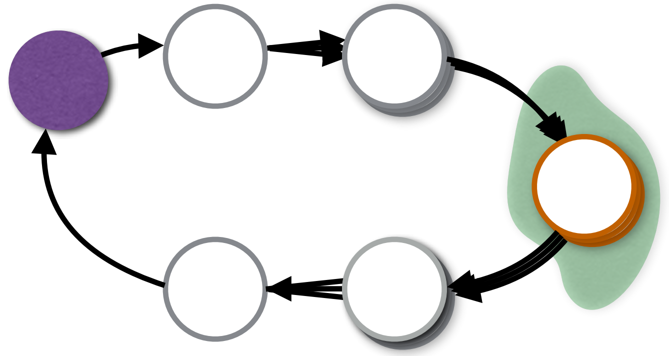
> "… tested thoroughly … written specifically to evade testing.
>
> But such evasion is likely to require some suspicious-looking code, …"

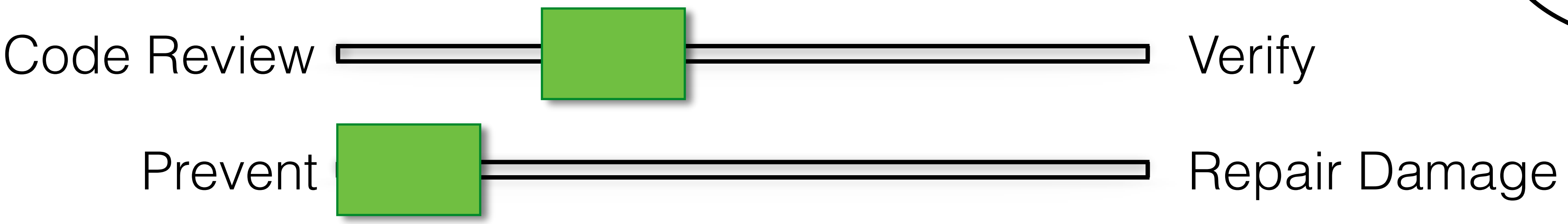
Building Reliable Voting Machine Software
Ka-Ping Yee 2007 dissertation

```
for (i = 0; i < limit; i++)
  if (j === 72374928)
```

Passes all tests.
Looks weird.

Inspect + Monitor





Code Review — Verify

Prevent — Repair Damage

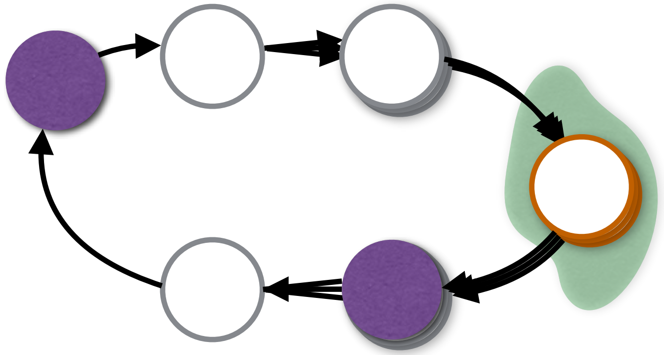
```
for (i = 0; i < limit; i++)
  if (j === 72374928)
```

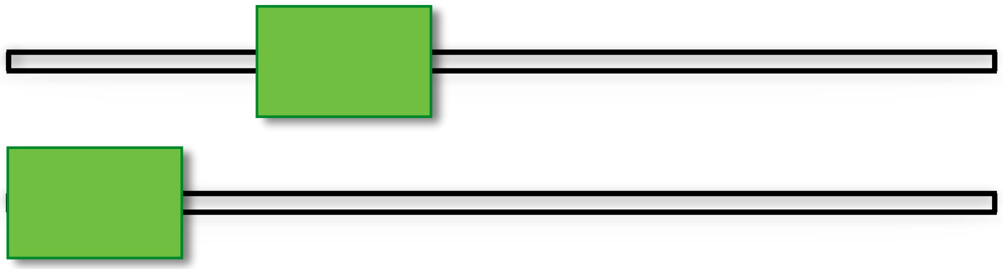
Passes all tests.
Looks weird.

```
for (i = 0; i <= limit; i++)
```

Looks fine.

Inspect + Monitor





```
for (i = 0; i < limit; i++)
  if (j === 72374928)
```

Passes all tests.
Looks weird.

```
for (i = 0; i <= limit; i++)
```

Fails on zero and one.
Looks fine.

rovi
rovi
SFO-SIDEREEL4-1
Bittman
How to Cook Everything
How to Cook Everything
Mark Bittman
SAMSUNG

Need Overall Conclusions

Get past the mutual disdain
Study the composition of compromises
Identify cross-bracing opportunities
Design for…
Design languages to support…
De-emphasize human vs object
  focus on cross cutting distinctions
  build mixed world
Incentives for people
Constraints for software

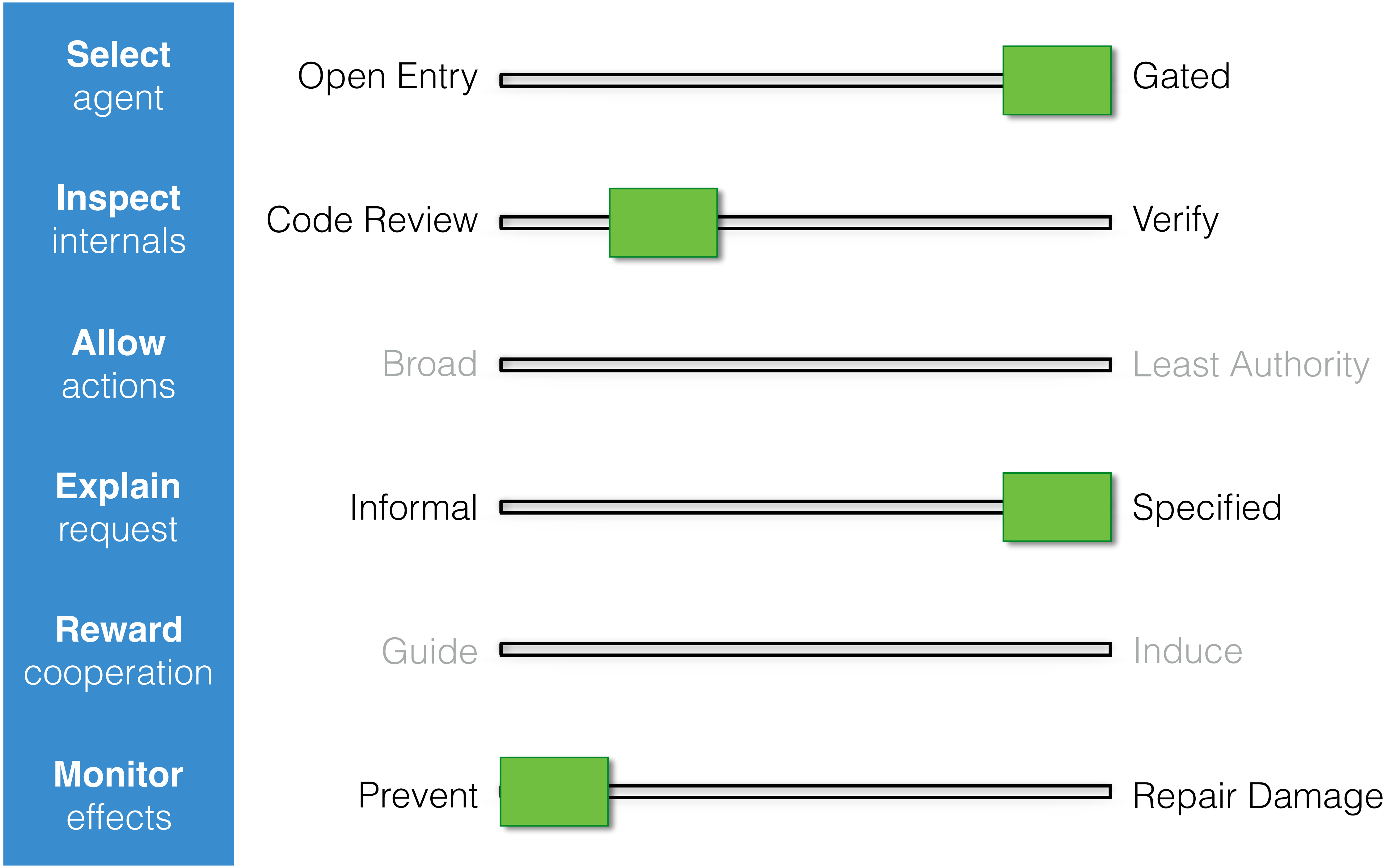
Tradeoff Alignments

	Concrete. Hierarchy	**Abstract. Decentralized**
Trust and Reputation	Admission controls. Aligned intentions	Open entry. Scalable
Static Inspection	Lint, code reviews. Find some bugs	Verify properties. Constrained behavior
Powers	Commons, administered. Low coordination costs	Narrow, transferable. Limit risk
Explanation	Informal understanding. Adaptive judgement	Explicit specification. Reuse, Competition
“Incentives”	Objective function. Inarticulate goals	Market prices. Aggregate tradeoffs
Dynamic Monitoring and Feedback	Defensive testing. Fail fast, report bugs	Intrusion detection. Repair damage

The Elements of Decision Alignment

	Human to Human	Human to Object	Object to Object
Select	Trademark Chain of custody	App stores White and black lists	Trusted developer Same origin
Inspect	Accounting controls	Trusted path URL bar	Types, Verification Open source eyeballs
Allow	Law, Contracts	App permissions Powerbox	Security Protection patterns
Explain	Language	User interface	Abstraction
Reward	Economics Incentive Alignment	Objective functions	Machine learning Agorics
Monitor	Reviews, Complaints Word of mouth	Bug reports	Contracts, Testing Backprop

Yee Voting Machine



Select agent — Open Entry — Gated

Inspect internals — Code Review — Verify

Allow actions — Broad — Least Authority

Explain request — Informal — Specified

Reward cooperation — Guide — Induce

Monitor effects — Prevent — Repair Damage