Distributed Meta-Management for Self-Protection and Self-Explanation

Catriona Kennedy

School of Computer Science,

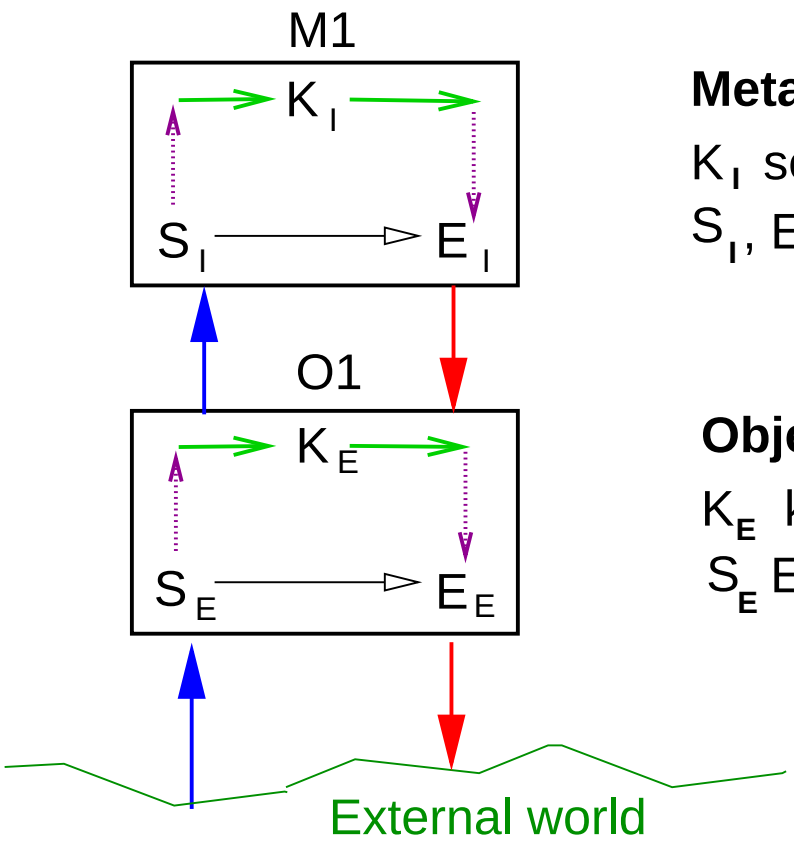
The University of Birmingham, UK

http://www.cs.bham.ac.uk/~cmk/

Overview

- **Part 1: Distributed Self-Protection**
  - Meta-management
  - Distributed and mutual meta-levels
  - Models of “self” in a distributed system
- **Part 2: From Self-Protection to Self-Explanation**
  - Examples of failure explanation
  - Comparison with human-like meta-cognition
  - Challenges: coordination

H−Cogaff Inspired Agent with Meta−level



Meta−level (M1)

K_I self−knowledge

S_I, E_I internal sensors/ effectors

Object−level (O1)

K_E knowledge of world

S_E E_E external sensors/ effectors

Perception and control

Reasoning

Reacting

Sensing and acting

Reactive + deliberative = object−level, External world = ground level,
Meta−management = meta−level

Meta-Management

- **Monitors** and critically evaluates **information states and processs** that the object level is **relying on;**
- **Runs concurrently**: but an unusual situation will increase meta-level processing.
- **Is context-aware**: information states **refer to things in the outside world**, not a purely syntactic trace.
- **Has a *model* of desirable and/or expected operation** of object-level: what should the trace look like, in what context? What does "failure" mean?

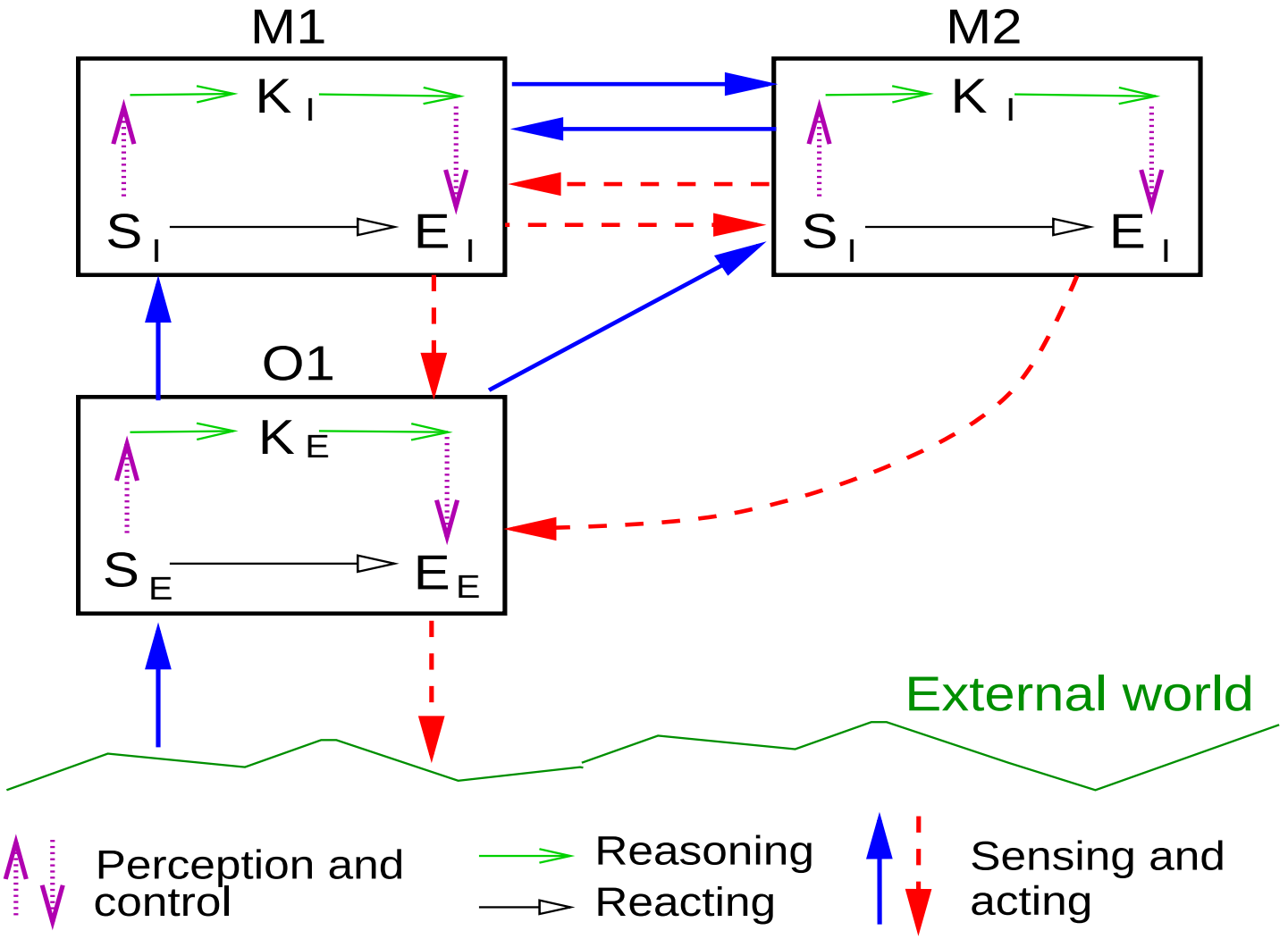
Uses of meta-management

- Detect failure or impending problem:
  - due to design fault (e.g. tendency to misinterpret a scene);
  - due to hostile attack, e.g. overload and diversion of computing resources, exploitation of design faults and vulnerabilities).
- Recognise own lack of knowlege:
  - after a failure has already occured
  - due to repeated differences between expectations (object-level knowledge) and reality.

Why Distributed Meta-management?

- Limits of centralised meta-management:
  - a meta-level may fail to detect a problem or even cause damage (e.g. divert resources, disrupt object-level processing).
  - its knowledge may be insufficient or incorrect.
  - a hostile environment can attack *any* part of the system, including the failure-detection and recovery components of the meta-level.
- Distribute meta-levels so that they mutually monitor and protect each other.
  - They also critically evaluate each other's reasoning and knowledge.
  - Monitoring is concurrent, but any intervention needs to be coordinated.

Distributed Meta−Levels − Configuration 1



Meta−level (M1, M2)

K_I self−knowledge

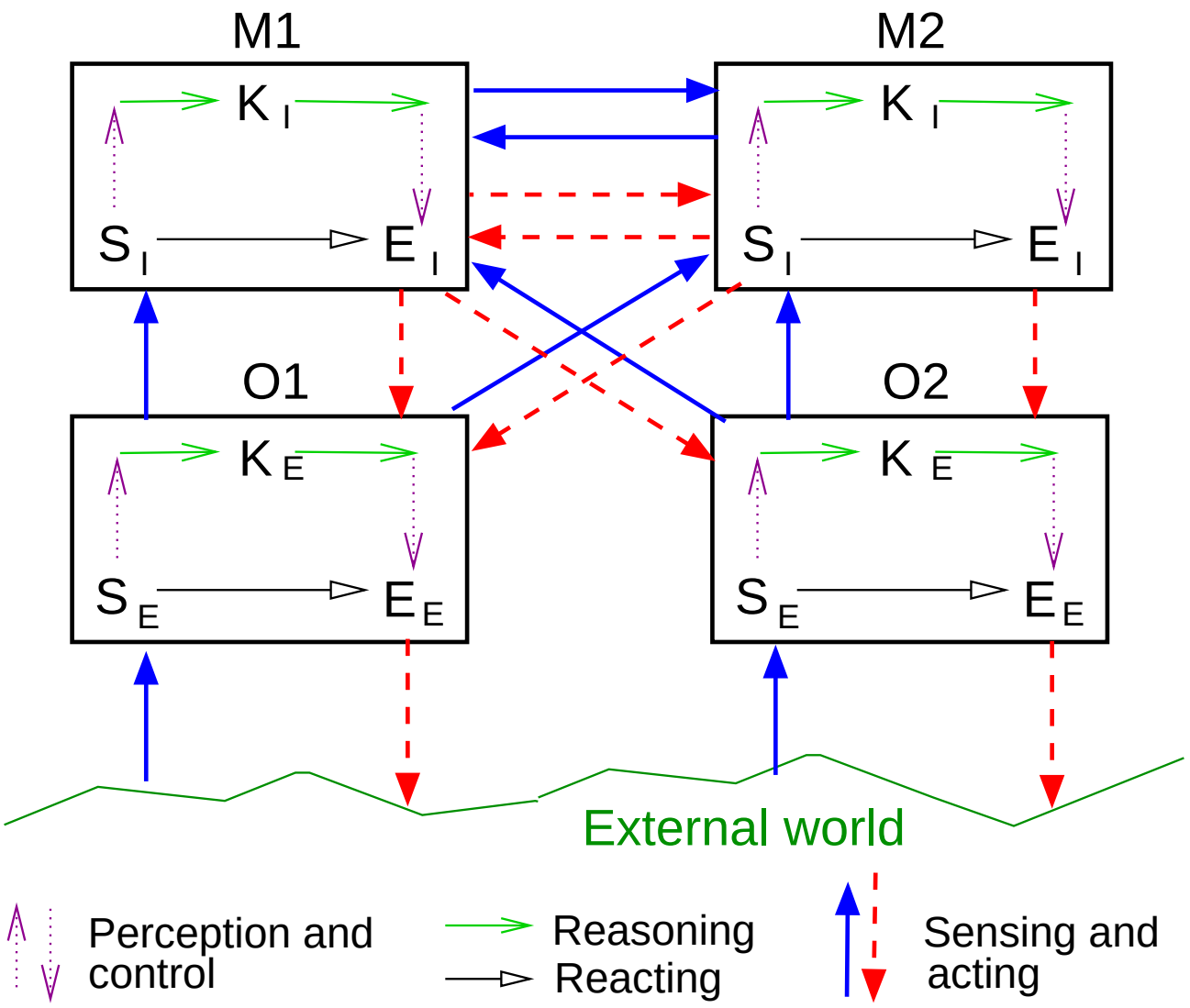
S_I, E_I internal sensors and effectors

Object−level (O1)

K_E Knowledge of world

S_E E_E external sensors and effectors

Distributed Meta−levels − Configuration 2



Meta−levels (M1, M2)

K_I self−knowledge

S_I, E_I internal sensors and effectors

Object−levels (O1, O2))

K_E knowledge of world

S_E E_E external sensors and effectors

Comparison with Raja and Lesser's Multi-Agent Meta-Levels

- **Differences:**
  - **mutual meta-level relationship: meta-levels are each other's "object-levels": no meta-levels should be un-monitored.**
  - **distributed meta-cognition for one agent - not separate individuals.**
- **Similarities:**
  - **many aspects of multi-agent coordination also apply to distributed meta-cognition.**
  - **although all meta-levels can intervene directly in each other's operation, it is often impractical to do so for many real world scenarios.**
  - **Ontologies of mental states and processes are required both for individual meta-cognition and for social meta-cognition (understanding of other minds).**

Towards "Closed" Meta-level Networks

- Aim: monitor and question *all* the major reasoning processes (including meta-level(s)).
- "Homunculus problem" is addressed, but new complexity is added (e.g. coordination and learning of self-models).
- Some gaps in coverage will still exist, but focus monitoring on most critical processes ("critical" is related to evolutionary fitness).
- No process that is important for agent survival should be un-monitored or unquestioned.
- Related concepts:
  - immune system models, Minsky's Society of Mind, Maturana and Varela's Organisational Closure and others in Second Order Cybernetics tradition.

Distributed Representations of "Self"

A self-representation includes:

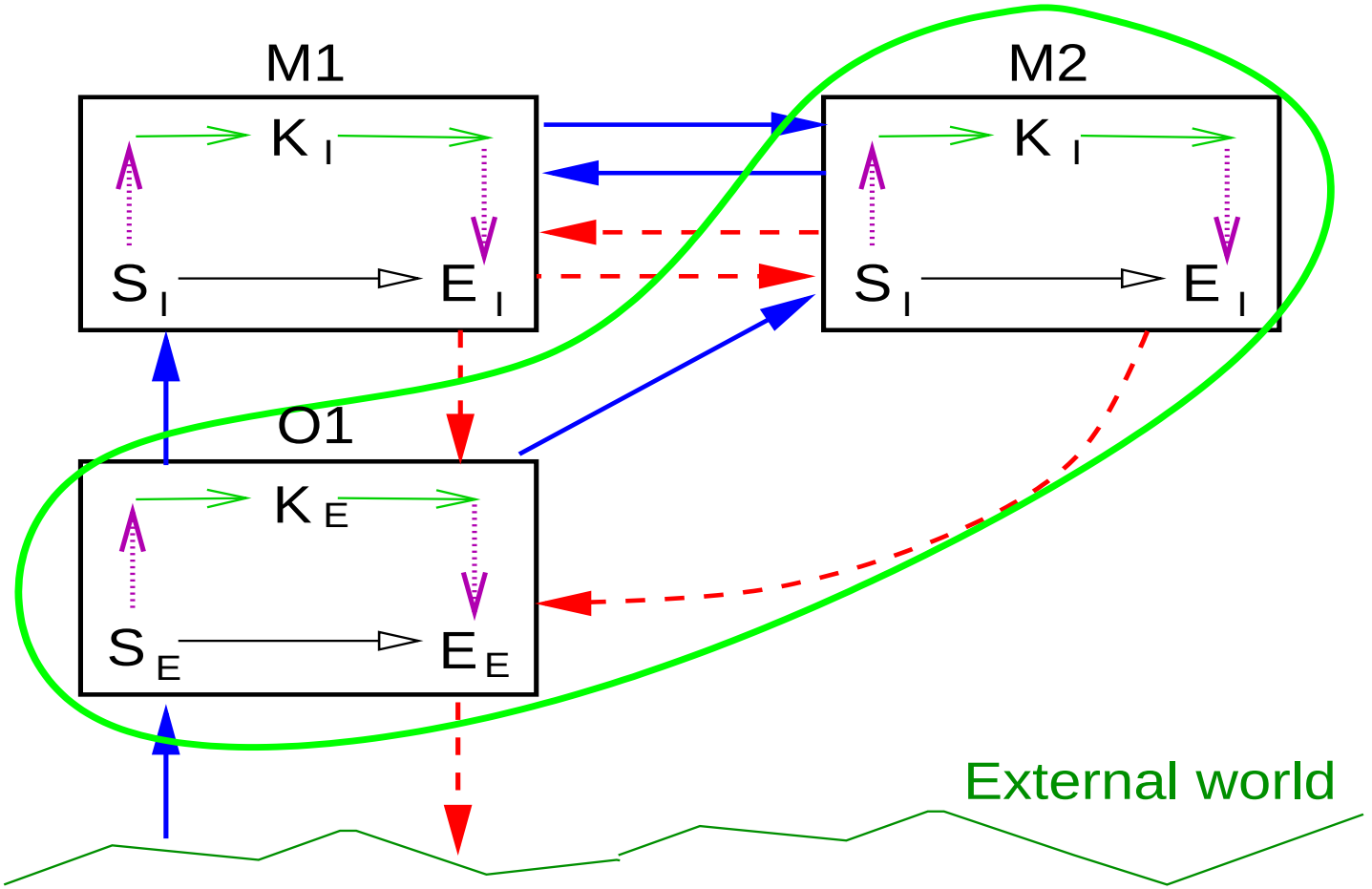
- **a map of components (e.g. collections of specialist rules).**
- **their current states along with recent history (e.g. rule firing trace).**
- **general patterns of normal operation aquired by self-familiarisation - immune system understanding of "self".**
- **general patterns of desirable operation (what methods and models have worked in the past?)**

Self-Familiarisation

Self-familiarisation by "mutual bootstrapping" of models (example for Configuration 1):

- **Phase 1: both meta-levels learn about normal object-level operation in protected enviroment.**
- **Phase 2: both meta-levels take turns observing each other responding to object-level failures.**
- **Note: very simplified description: many iterations of mutual observation and learning are possible.**
- **Should really take place in in parallel in a variety of simulated hostile environments.**
- **Not just one self-representation - multiple partial views.**

Self−Representation − M1's view



Perception and control

Reasoning

Reacting

Sensing and acting

Meta−level (M1, M2)

K_I self−knowledge

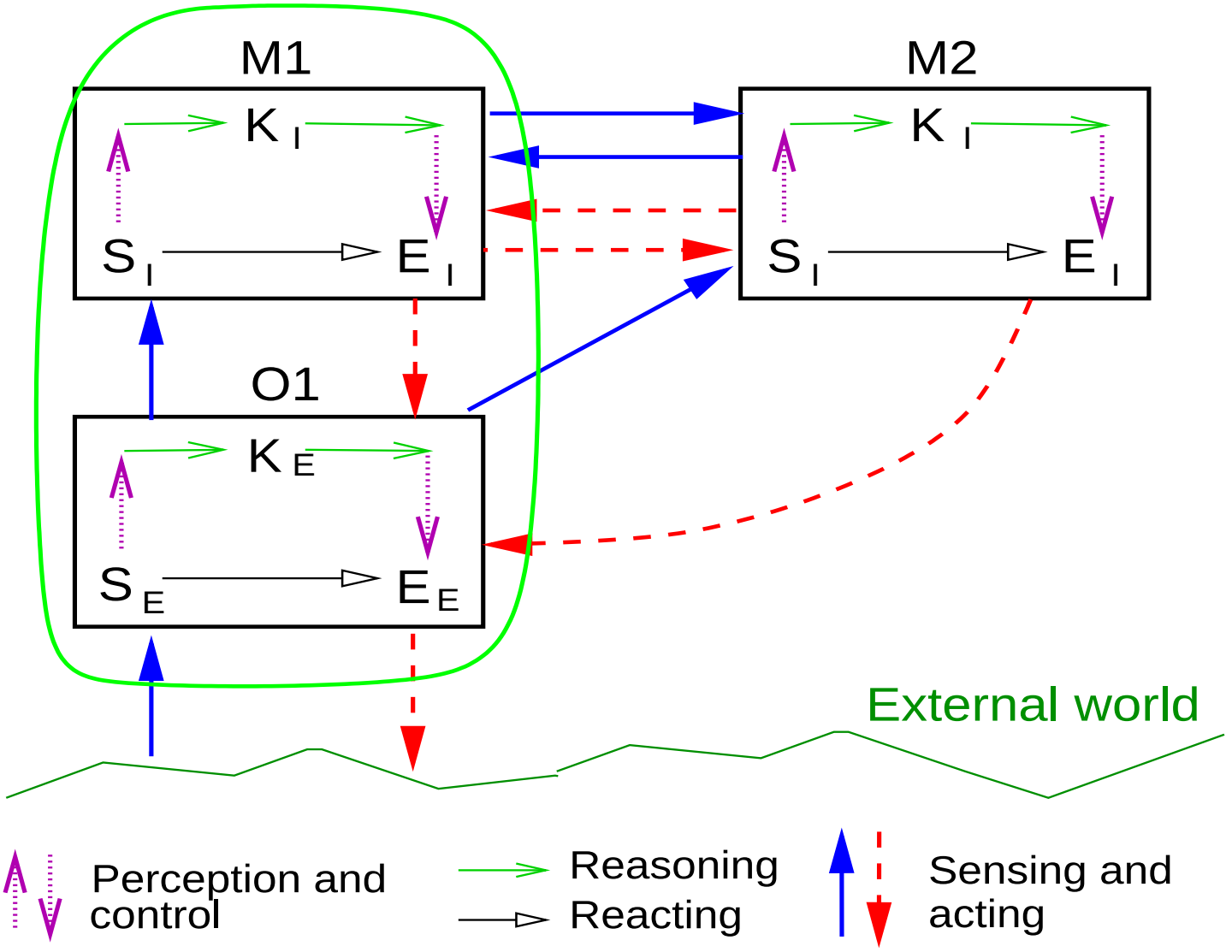
S_I, E_I internal sensors and effectors

Object−level (O1)

K_E Knowledge of world

S_E E_E external sensors and effectors

Self−representation − M2's view



Perception and control

Reasoning

Reacting

Sensing and acting

Meta−level (M1, M2)

K_I self−knowledge

S_I, E_I internal sensors and effectors

Object−level (O1)

K_E Knowledge of world

S_E E_E external sensors and effectors

From Self-Protection to Self-Explanation

- Self-explanation is important for any meta-reasoning system (Cox, 2007).
- In a self-protection context, important to understand and explain *why* a meta-level is failing.
- In a distributed system, coordination is required to produce a single coherent narrative.
- How to translate from objective software interactions to human-like meta-cognition?
- Link together competing hypotheses about cause of failure.

Distributed Self-Explanation: Example 1

Example 1: competing hypotheses about an object-level failure:

- **Two meta-levels M1 and M2 are monitoring the same object-level O1, but in different ways.**
- **M1, M2 and O1 all produce a reasoning trace.**
- **M1's hypothesis: "O1's reasoning at step S was incorrect";**
- **M2's hypothesis: "M1 wrongly detected an error in O1 at step S; the problem is in step U;"**
- **M2 may also explain why M1 is incorrect, also by analysing its trace.**

Translation into Human-like Meta-Cognition

The competing hypotheses may be translated into a human-like explanation as follows:

"I suspect I made a mistake at step S but I can't rule out other possibilities, maybe step S was correct but step U was wrong."

Part of sentence	Meaning of "I"
"I suspect that .."	M_1
"I made a mistake at step S .."	O_1
"but I can't rule out other possibilities ..."	M_2

Example 2: Lack of Knowledge

A meta-level M2 reasons about another meta-level M1's knowledge:

- **Both meta-levels are monitoring the same object-level O1, as in the previous example.**
- **M1's hypothesis: "O1's knowledge about concept C is incorrect";**
- **M2's hypothesis: "M1 is wrong about O1's knowledge, the problem is elsewhere";**
- **M2 may also explain that M1's knowledge is insufficient for this particular problem, e.g. because of the way it is interpreting O1's trace.**

Challenge: Coordination

Disagreements between meta-levels - how to prevent oscillations:

- **first meta-level to make a decision can inhibit all others.**
- **it should be given sufficient time to attempt to solve the problem.**
- **if it is not making any progress, another meta-level can interrupt it.**
- **more complex coordination required in a hostile environment: e.g. agreement among a majority of meta-levels before one is allowed to proceed.**
- **oscillations may still take place over extended time intervals - but this also true of humans.**

Challenge: Global Explanation

How to ensure a sequential narrative:

- **the first meta-level to construct an atomic part of an explanation (e.g. a sentence) can broadcast its belief and inhibit all others - in the same way as for action coordination.**
  - **– remaining meta-levels then state their beliefs.**
- **Alternatively, a single meta-level may collect belief states of all others and chain them together into a single statement.**
- **Both kinds of meta-level can have their explanation interrupted.**

Challenge: Cost of Meta-Management

- For a single meta-level, cost may be measured in several ways:
  - how much additional computing time is required?
  - how much does the meta-level reasoning interfere with the problem solving on the object-level? (e.g. incorrect interventions).
- Such costs can be monitored by other meta-levels in a distributed system.
- Additional costs arise from distribution:
  - **Communication and connectivity:** Can be mitigated if system can evolve into a state in which only useful connections are preserved, and where meta-level overheads are counteracted by increased reliability. Distributed agreement would be required before severing a connection.
  - **Self-familiarisation adds complexity:** Distributed meta-levels require coordination, and the system needs to be familiar with the coordinating mechanisms, adding yet more complexity required to bootstrap a self-representation. However, software component re-use can avoid the need for more components.

Summary and Conclusions

- For autonomic computing and robotics, self-protection against faults and intrusions requires non-hierarchical architecture.
- However, self-explanation is also important, not only for the system itself but also to enable humans interacting with it to understand the reasons for its actions.
- Reconciling these two requirements is possible, but requires an integrated, cross-disciplinary approach to cognitive systems.
- In addition to AI methods, research in distributed fault-tolerance, software engineering and cognitive neuroscience can make a valuable contribution.