2. Responsibility:

Designers and builders of advanced AI systems are stakeholders in the moral implications of their use, misuse, and actions, with a responsibility and opportunity to shape those implications.
2. Responsibility:

Designers and builders of advanced AI systems are stakeholders in the moral implications of their use, misuse, and actions, with a responsibility and opportunity to shape those implications.

- A distinction should be made between those who design/build the initial AI systems and those who teach the AI systems.
- Glosses over the (life-long) learning of the AI.
- Teachers bear the greater responsibility, and so do the people / institutions that accredit, manage, and monitor those AI teachers.
10. Recursive self-improvement:

AI systems should be designed and created primarily by humans: an AI system that creates or modifies algorithms, including its own, must do so in a way that retains verifiable safety of the full new system.
10. Recursive self-improvement:

AI systems should be designed and created primarily by humans: an AI system that creates or modifies algorithms, including its own, must do so in a way that retains verifiable safety of the full new system.

- The described approach to RSI is wrong and unsafe
- New: “AI systems designed to self-improve or self-replicate in a manner that could lead to exponentially increasing quality or quantity must be subject to strict safety and control measures.”
- Still fails to put the finger at the crux of the matter—the crux is in getting the AI to understand how chains of actions and events lead to the violation of ethical constraints
11A. Existential risk:

No AI system should be created with a conceivable chance of representing a global catastrophic or existential risk unless credible disinterested expert analysis shows the risk to be worth taking.
### 11A. Existential risk:

No AI system should be created with a conceivable chance of representing a global catastrophic or existential risk unless credible disinterested expert *analysis* shows the risk to be worth taking.

- The failure to acknowledge the centrality of interactive *teaching & testing* leads to a black-box-behaviorist way of thinking
- Must focus on the *process*, not the *result*
- Otherwise we risk a fear-induced, after-the-fact, symptom-fighting scramble for control
How to approach Recursive Self-Improvement

Typical question

“What is the behavior of an AI that is very intelligent and capable of self-modification—and how do we control it?”
How to approach Recursive Self-Improvement

Wrong question

“What is the behavior of an AI that is very intelligent and capable of self-modification—and how do we control it?”

Right question

“How do we grow an AI from baby beginnings such that it gains both robust understanding and proper ethics?”
Ultimate aim (of AI safety)

We want AIs to be *compelled* to adhere to *ethical values*, throughout their lifetimes, despite possible *interference* and recursive self-improvement.
Ultimate aim (of AI safety)

We want AIs to be compelled to adhere to ethical values, throughout their lifetimes, despite possible interference and recursive self-improvement.

Ethics > understanding

Adherence to ethics requires understanding of how chains of actions and events lead to the violation thereof.
We humans want AI with *bounded* recursive self-improvement.

1. Bounded by tasks (requirements to meet, constraints to respect)
2. Bounded by ethics (across tasks and independent thereof)
3. Bounded by resource and knowledge limitations
We humans want AI with *bounded* recursive self-improvement.

1. Bounded by tasks (requirements to meet, constraints to respect)
2. Bounded by ethics (across tasks and independent thereof)
3. Bounded by resource and knowledge limitations

These bounds may be unknown beforehand and changing over time!
Can we now be more specific about what RSI must do in order to allow for progress of understanding and adherence to ethics?
Can we now be more specific about what RSI must do in order to allow for progress of understanding and adherence to ethics?

- We must aim to identify and qualify the internal constituent components that give rise to understanding.
- Therefore we must specify what constitutes “self-modification,” so that we can tell whether or not a particular self-modification is in service of making progress in understanding.
- …of ethical constraints, especially.
- Conduct many pressure tests over time, to grow and test understanding of (ethical) constraints.
Experience-based AI (EXPAI)

The Idea

Self-modifications shall be FATRR

1. Fine-grained
2. Additive
3. Tentative
4. Rated over time
5. Revertible
Experience-based AI (EXPAI)

The Idea

Self-modifications shall be FATRR

1. Fine-grained
2. Additive
3. Tentative
4. Rated over time
5. Revertible

Implications

- No reasoning or proofs about self-modifications needed
- Experience-based vindication & falsification
- Backward-looking “proof” of self-improvements
Knowledge represented as *granules*, grown from a “seed”

Functionality of forward and inverse models (Control Theory)

Allow chaining (horizontal) and hierarchy (vertical)

*Requirements* specifiable as goals and constraints

Knowledge *decoupled* from goals

Controller *dynamically couples* knowledge & goals → actions

*Simulation* before commitment
Experience-based AI
Example demonstrator

Not hot air: ≥ 1 implementation exists

Autocatalytic Endogenous Reflective Architecture (AERA)
In order to control a powerful entity, the controlling entity must be at least as powerful.

For AIs that can grow to become significantly more powerful than humans (and their tools), the only way to control them is for them to control themselves.
In order to **control** a powerful entity, the controlling entity must be at least as powerful.

For AIs that can grow to become significantly more powerful than humans (and their tools), the only way to control them is for them to **control themselves**.

**Corollary**

Ethical (meta-)values are constraints that must **stabilize** over time.

- Otherwise we have no assurances about the long-term self-constrained behavior of the AI.
- Stabilization must occur **before** the AI becomes too powerful to control it directly (before it’s capable of preventing someone—physically or persuasively—from pressing the off-switch).
Closing Thoughts

For AI safety, we need to get three things right

1. The architecture of the AI at start-up
2. The teaching of the AI, to develop the understanding of (ethical) constraints
3. “Complete” the teaching before the “deadline of control”

AI may be “softer” than it’s been so far

More responsibility on teachers than programmers
YOU ENCOUNTER IT WHEN YOU FIRST STUDY PHYSICS. YOU REALIZE THAT, IF YOU WERE EVER DROPPED FROM A PLANE WITHOUT A PARACHUTE, YOU COULD CALCULATE WITH A HIGH DEGREE OF ACCURACY HOW LONG IT'D TAKE TO HIT THE GROUND, YOUR SPEED, HOW MUCH ENERGY YOU'LL DEPOSIT INTO THE EARTH.

AND YET, YOU WOULD STILL BE JUST AS DEAD AS A PARTICULARLY STUPID GORILLA DROPPED THE SAME DISTANCE.

I COULD TELL YOU WHY GRANDPA IS VERY SICK. I COULD TELL YOU WHAT EACH CELL IS DOING WRONG. WHY IT'S DOING WRONG, AND ROUGHLY WHEN IT STARTED DOING WRONG.

BUT I CAN'T TELL THEM TO STOP.

I CALL THAT THE FALLING PROBLEM.

MASTERY OF THE NATURE OF REALITY GRANTS YOU NO MASTERY OVER THE BEHAVIOR OF REALITY.

WHY CAN'T YOU MAKE A MACHINE TO FIX IT?

SAME REASON YOU CAN'T MAKE A PARACHUTE WHEN YOU FALL FROM THE PLANE.

BECauses IT'S TOO HARD?

NOTHING IS TOO HARD. MANY THINGS ARE TOO FAST.

I THINK I COULD SOLVE THE FALLING PROBLEM WITH A JETPACK. CAN YOU TRY TO GET ME THE PARTS?

THAT'S ALL I DO, KIDDO.