An overview of *Superintelligence*, by Nick Bostrom

Alistair Knott

The unfinished fable of the sparrows

The unfinished fable of the sparrows

It was the nest-building season, but after days of long hard work, the sparrows sat in the evening glow, relaxing and chirping away.
“We are all so small and weak. Imagine how easy life would be if we had an owl who could help us build our nests!”

The unfinished fable of the sparrows

It was the nest-building season, but after days of long hard work, the sparrows sat in the evening glow, relaxing and chirping away.
“We are all so small and weak. Imagine how easy life would be if we had an owl who could help us build our nests!”

Pastus, the elder-bird, spoke: “Let us send out scouts in all directions and try to find an abandoned owlet somewhere, or maybe an egg.”

The unfinished fable of the sparrows

Only Scronkfinkle was unconvinced. Quoth he: “This will surely be our undoing. Should we not give some thought to the art of owl-domestication and owl-taming first, before we bring such a creature into our midst?”

The unfinished fable of the sparrows

Only Scronkfinkle was unconvinced. Quoth he: “This will surely be our undoing. Should we not give some thought to the art of owl-domestication and owl-taming first, before we bring such a creature into our midst?”

Replied Pastus: “Taming an owl sounds like an exceedingly difficult thing to do. It will be difficult enough to find an owl egg. So let us start there. After we have succeeded in raising an owl, then we can think about taking on this other challenge.”

The unfinished fable of the sparrows

Only Scronkfinkle was unconvinced. Quoth he: “This will surely be our undoing. Should we not give some thought to the art of owl-domestication and owl-taming first, before we bring such a creature into our midst?”

Replied Pastus: “Taming an owl sounds like an exceedingly difficult thing to do. It will be difficult enough to find an owl egg. So let us start there. After we have succeeded in raising an owl, then we can think about taking on this other challenge.”

“There is a flaw in that plan!” squeaked Scronkfinkle; but his protests were in vain. . .

Structure of the book

Structure of the book

- Forms of superintelligence
- The kinetics of an intelligence explosion
- The powers of a superintelligent agent
- The superintelligent will
- The control problem
- Acquiring values

1. Forms of superintelligence

- Speed superintelligence
- Collective superintelligence
- Quality superintelligence

1. Forms of superintelligence

- Speed superintelligence
- Collective superintelligence
- Quality superintelligence

Advantages of having intelligence implemented in software:

- Editability
- Duplicability
- Memory sharing
- Module-based design

2. The kinetics of an intelligence explosion

2. The kinetics of an intelligence explosion

Rate of change of a machine's intelligence = $\frac{\text{optimisation power}}{\text{recalcitrance}}$

2. The kinetics of an intelligence explosion

Rate of change of a machine's intelligence = $\frac{\text{optimisation power}}{\text{recalcitrance}}$

- Optimisation power: the capability (of humans and/or the machine) to improve the machine.

2. The kinetics of an intelligence explosion

Rate of change of a machine's intelligence = $\frac{\text{optimisation power}}{\text{recalcitrance}}$

- Optimisation power: the capability (of humans and/or the machine) to improve the machine.
- Recalcitrance: the difficulty of improving the machine.

2. The kinetics of an intelligence explosion

Things that would increase optimisation power:

2. The kinetics of an intelligence explosion

Things that would increase optimisation power:

- If one AI paradigm shows promise, other human AI researchers will pile in to extend/improve it.

2. The kinetics of an intelligence explosion

Things that would increase optimisation power:

- If one AI paradigm shows promise, other human AI researchers will pile in to extend/improve it.
- If the AI system gets good enough, it can improve its *own* design. (Bostrom calls that ‘crossover’.)

2. The kinetics of an intelligence explosion

Things that would increase optimisation power:

- If one AI paradigm shows promise, other human AI researchers will pile in to extend/improve it.
- If the AI system gets good enough, it can improve its *own* design. (Bostrom calls that ‘crossover’.)
  Improvements here could have an exponential character.

2. The kinetics of an intelligence explosion

Things that could reduce recalcitrance:

2. The kinetics of an intelligence explosion

Things that could reduce recalcitrance:

- The discovery of ‘one key insight’ that has been holding things back.

2. The kinetics of an intelligence explosion

Things that could reduce recalcitrance:

- The discovery of ‘one key insight’ that has been holding things back.
- The addition of mechanisms that allow the machine to learn from existing knowledge sources (e.g. the Library of Congress).

2. The kinetics of an intelligence explosion

Things that could reduce recalcitrance:

- The discovery of ‘one key insight’ that has been holding things back.
- The addition of mechanisms that allow the machine to learn from existing knowledge sources (e.g. the Library of Congress).
- Increases in computing power.

2. The kinetics of an intelligence explosion

Things that could reduce recalcitrance:

- The discovery of ‘one key insight’ that has been holding things back.
- The addition of mechanisms that allow the machine to learn from existing knowledge sources (e.g. the Library of Congress).
- Increases in computing power.

Our anthropocentric view of intelligence may lead us to overestimate recalcitrance.

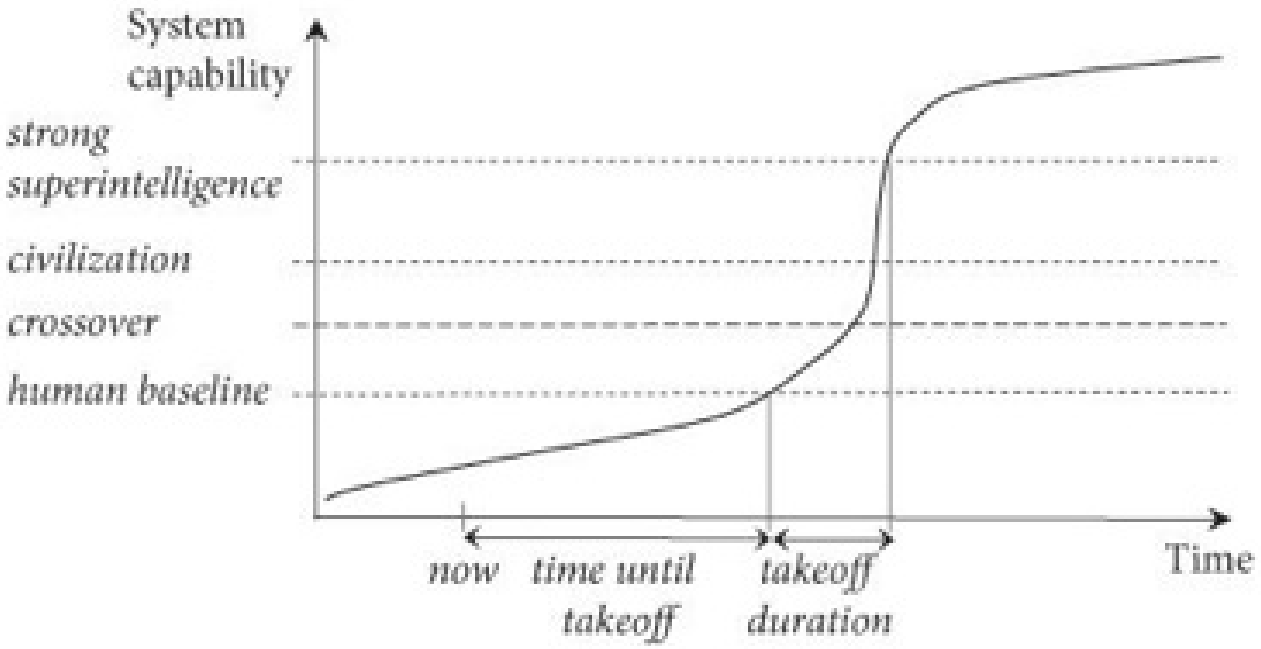
2. The kinetics of an intelligence explosion

Things that could reduce recalcitrance:

- The discovery of ‘one key insight’ that has been holding things back.
- The addition of mechanisms that allow the machine to learn from existing knowledge sources (e.g. the Library of Congress).
- Increases in computing power.

Our anthropocentric view of intelligence may lead us to overestimate recalcitrance.

2. The kinetics of an intelligence explosion



3. The powers of a superintelligent agent

3. The powers of a superintelligent agent

A superintelligent agent would have several abilities:

3. The powers of a superintelligent agent

A superintelligent agent would have several abilities:

- *Strategic abilities* (to achieve goals, overcome intelligentopposition)

3. The powers of a superintelligent agent

A superintelligent agent would have several abilities:

- *Strategic abilities* (to achieve goals, overcome intelligentopposition)
- *Social abilities* (to manipulate people into doing what it wants)

3. The powers of a superintelligent agent

A superintelligent agent would have several abilities:

- *Strategic abilities* (to achieve goals, overcome intelligentopposition)
- *Social abilities* (to manipulate people into doing what it wants)
- *Economic abilities* (abilities to make lots of money)

3. The powers of a superintelligent agent

A superintelligent agent would have several abilities:

- *Strategic abilities* (to achieve goals, overcome intelligentopposition)
- *Social abilities* (to manipulate people into doing what it wants)
- *Economic abilities* (abilities to make lots of money)
- *Technical abilities* (abilities to invent/build machines)

3. The powers of a superintelligent agent

A superintelligent agent would have several abilities:

- *Strategic abilities* (to achieve goals, overcome intelligentopposition)
- *Social abilities* (to manipulate people into doing what it wants)
- *Economic abilities* (abilities to make lots of money)
- *Technical abilities* (abilities to invent/build machines)
- *Hacking abilities* (e.g. to find holes in security systems).

4. The superintelligent will

4. The superintelligent will

‘We have already cautioned against anthropomorphising the *capabilities* of a superintelligent AI. This warning should be extended to pertain to its *motivations* as well.’

4. The superintelligent will

‘We have already cautioned against anthropomorphising the *capabilities* of a superintelligent AI. This warning should be extended to pertain to its *motivations* as well.’

- The orthogonality thesis
  Intelligence and final goals are orthogonal: more or less any intelligence could in principle be combined with more or less any final goal.

4. The superintelligent will

‘We have already cautioned against anthropomorphising the *capabilities* of a superintelligent AI. This warning should be extended to pertain to its *motivations* as well.’

- The orthogonality thesis
  Intelligence and final goals are orthogonal: more or less any intelligence could in principle be combined with more or less any final goal.

‘There is nothing paradoxical about an AI whose sole final goal is. . . to calculate the decimal expansion of pi, or to maximise the total number of paperclips in its future light cone.’

4. The superintelligent will

It may still be possible to make *predictions* about the motivation of a superintelligent machine.

4. The superintelligent will

It may still be possible to make *predictions* about the motivation of a superintelligent machine.

Perhaps there are certain ‘instrumental’ goals that any superintelligent agent would adopt to further its ultimate goal.

4. The superintelligent will

It may still be possible to make *predictions* about the motivation of a superintelligent machine.

Perhaps there are certain ‘instrumental’ goals that any superintelligent agent would adopt to further its ultimate goal.

- Self-preservation
- Retention of goals through time
- Cognitive enhancement
- Technological perfection
- Resource acquisition

5. The control problem

5. The control problem

If we can't *control* the superintelligent agent, the result will likely be catastrophic. (For us.)

5. The control problem

If we can't *control* the superintelligent agent, the result will likely be catastrophic. (For us.)

There are two ways we might control it:

5. The control problem

If we can't *control* the superintelligent agent, the result will likely be catastrophic. (For us.)

There are two ways we might control it:

- Capability control (limiting what the system *can* or *does* do).

5. The control problem

If we can't *control* the superintelligent agent, the result will likely be catastrophic. (For us.)

There are two ways we might control it:

- Capability control (limiting what the system *can* or *does* do).
- Motivation selection (controlling what the system *wants* to do).

5. The control problem

Capability control methods:

5. The control problem

Capability control methods:

- *Boxing*: the system can only act through restricted channels.

5. The control problem

Capability control methods:

- *Boxing*: the system can only act through restricted channels.
- *Incentives*: access to other AIs, cryptographic reward tokens...

5. The control problem

Capability control methods:

- *Boxing*: the system can only act through restricted channels.
- *Incentives*: access to other AIs, cryptographic reward tokens...
- *Stunting*: imposing constraints on the system's cognitive abilities

5. The control problem

Capability control methods:

- *Boxing*: the system can only act through restricted channels.
- *Incentives*: access to other AIs, cryptographic reward tokens...
- *Stunting*: imposing constraints on the system's cognitive abilities
- *Tripwires*: diagnostic tests run periodically to check for dangerous activity, with shutdown a consequence of detection.

5. The control problem

Capability control methods:

- *Boxing*: the system can only act through restricted channels.
- *Incentives*: access to other AIs, cryptographic reward tokens...
- *Stunting*: imposing constraints on the system's cognitive abilities
- *Tripwires*: diagnostic tests run periodically to check for dangerous activity, with shutdown a consequence of detection.

I'll focus on motivation selection methods. (How might we control what the system *wants* to do?)

(i) Direct specification of motivations

(i) Direct specification of motivations

Rule-based methods: give the machine a set of rules that define its final goals.

(i) Direct specification of motivations

Rule-based methods: give the machine a set of rules that define its final goals.

- But: it's hard/impossible to specify a set of rules that is *precise*/*consistent*. (As lawyers know.)

(i) Direct specification of motivations

Rule-based methods: give the machine a set of rules that define its final goals.

- But: it's hard/impossible to specify a set of rules that is *precise*/*consistent*. (As lawyers know.)

Direct consequentialist methods: specify some measure that is to be maximised. (E.g. human happiness.)

(i) Direct specification of motivations

Rule-based methods: give the machine a set of rules that define its final goals.

- But: it's hard/impossible to specify a set of rules that is *precise*/*consistent*. (As lawyers know.)

Direct consequentialist methods: specify some measure that is to be maximised. (E.g. human happiness.)

- But: these could be interpreted in ways we didn't foresee.

(ii) Augmentation

(ii) Augmentation

Start with an AI with human-level intelligence, that has an acceptable motivation system: then enhance its cognitive faculties to make it superintelligent.

(ii) Augmentation

Start with an AI with human-level intelligence, that has an acceptable motivation system: then enhance its cognitive faculties to make it superintelligent.

- ‘If all goes well, this would give us a superintelligence with an acceptable motivation system.’

(ii) Augmentation

Start with an AI with human-level intelligence, that has an acceptable motivation system: then enhance its cognitive faculties to make it superintelligent.

- ‘If all goes well, this would give us a superintelligence with an acceptable motivation system.’

But:

(ii) Augmentation

Start with an AI with human-level intelligence, that has an acceptable motivation system: then enhance its cognitive faculties to make it superintelligent.

- ‘If all goes well, this would give us a superintelligence with an acceptable motivation system.’

But:

- This only works if we can start with a human-like AI.

(ii) Augmentation

Start with an AI with human-level intelligence, that has an acceptable motivation system: then enhance its cognitive faculties to make it superintelligent.

- ‘If all goes well, this would give us a superintelligence with an acceptable motivation system.’

But:

- This only works if we can start with a human-like AI.
- Even then, a humanlike intelligence might get corrupted in unpredictable ways in the ‘enhancement’ process.

(iii) Indirect normativity

(iii) Indirect normativity

‘Rather than specifying a normative standard directly, we specify a process for *deriving* a standard. We then build the system so it is motivated to carry out this process.’

(iii) Indirect normativity

‘Rather than specifying a normative standard directly, we specify a process for *deriving* a standard. We then build the system so it is motivated to carry out this process.’

- An example: ‘achieve that which we would have wished you [the AI] to achieve if we [humans] had thought about the matter long and hard’.

(iii) Indirect normativity

‘Rather than specifying a normative standard directly, we specify a process for *deriving* a standard. We then build the system so it is motivated to carry out this process.’

- An example: ‘achieve that which we would have wished you [the AI] to achieve if we [humans] had thought about the matter long and hard’.

Yudkowsky: a seed AI should be given the goal of carrying out humanity’s coherent extrapolated volition, defined as ‘our wish if we knew more, thought faster, were more the people we wished we were, had grown up farther together; where the extrapolation converges rather than diverges, where our wishes cohere rather than interfere; extrapolated as we wish that extrapolated, interpreted as we wish that interpreted’.

6. Acquiring values

6. Acquiring values

The indirect normativity strategy faces a technical problem, the value-loading problem.

6. Acquiring values

The indirect normativity strategy faces a technical problem, the value-loading problem.

- An *unintelligent agent* wouldn't be able to understand a complex indirect goal specification, let alone implement it.

6. Acquiring values

The indirect normativity strategy faces a technical problem, the value-loading problem.

- An *unintelligent agent* wouldn't be able to understand a complex indirect goal specification, let alone implement it.
- So you have to give it a simpler goal.

6. Acquiring values

The indirect normativity strategy faces a technical problem, the value-loading problem.

- An *unintelligent agent* wouldn't be able to understand a complex indirect goal specification, let alone implement it.
- So you have to give it a simpler goal.
- So how do you guarantee it *ends up* with the right complex goal when it gets more intelligent?

6. Acquiring values

The indirect normativity strategy faces a technical problem, the value-loading problem.

- An *unintelligent agent* wouldn't be able to understand a complex indirect goal specification, let alone implement it.
- So you have to give it a simpler goal.
- So how do you guarantee it *ends up* with the right complex goal when it gets more intelligent?

Bostrom has a few ideas. . .

(i) Associative value accretion

(i) Associative value accretion

This scheme begins from the observation that *we humans* must somehow acquire our values: we're not born with them.

(i) Associative value accretion

This scheme begins from the observation that *we humans* must somehow acquire our values: we're not born with them.

- We may acquire a desire for the wellbeing of some person *P*: but we don't start off with the concepts 'wellbeing' and 'P'.

(i) Associative value accretion

This scheme begins from the observation that *we humans* must somehow acquire our values: we're not born with them.

- We may acquire a desire for the wellbeing of some person *P*: but we don't start off with the concepts 'wellbeing' and 'P'.

So we could try modelling the process by which we acquire our concepts, and refine our values accordingly.

(i) Associative value accretion

This scheme begins from the observation that *we humans* must somehow acquire our values: we're not born with them.

- We may acquire a desire for the wellbeing of some person *P*: but we don't start off with the concepts 'wellbeing' and 'P'.

So we could try modelling the process by which we acquire our concepts, and refine our values accordingly.

- But: we don't know how this works yet—so we don't know what the pitfalls are.

(ii) Value learning

(ii) Value learning

In this scheme, the AI's final goal is specified as a *learning* goal, to *learn* the (first-order) values it adheres to in its behaviour.

(ii) Value learning

In this scheme, the AI's final goal is specified as a *learning* goal, to *learn* the (first-order) values it adheres to in its behaviour.

- It will begin with an imperfect (but usable) *approximation* of the first-order human values we want it to have.

(ii) Value learning

In this scheme, the AI's final goal is specified as a *learning* goal, to *learn* the (first-order) values it adheres to in its behaviour.

- It will begin with an imperfect (but usable) *approximation* of the first-order human values we want it to have.
- This approximation will become better as it learns.

(ii) Value learning

What does the AI learn human values *from*?

(ii) Value learning

What does the AI learn human values *from*?

- The basic idea is it should learn from observing human behaviour, as broadly as possible.

(ii) Value learning

What does the AI learn human values *from*?

- The basic idea is it should learn from observing human behaviour, as broadly as possible.
- Bostrom suggests it should try and *guess* what values the designers have specified for it. It will compute a distribution over many alternative hypotheses, and act accordingly.

(ii) Value learning

What does the AI learn human values *from*?

- The basic idea is it should learn from observing human behaviour, as broadly as possible.
- Bostrom suggests it should try and *guess* what values the designers have specified for it. It will compute a distribution over many alternative hypotheses, and act accordingly.

Aside: Stuart Russell suggests inverse reinforcement learning.

(ii) Value learning

What does the AI learn human values *from*?

- The basic idea is it should learn from observing human behaviour, as broadly as possible.
- Bostrom suggests it should try and *guess* what values the designers have specified for it. It will compute a distribution over many alternative hypotheses, and act accordingly.

Aside: Stuart Russell suggests inverse reinforcement learning.

- Model human agents as having a reward function, and acting so as to maximise reward.

(ii) Value learning

What does the AI learn human values *from*?

- The basic idea is it should learn from observing human behaviour, as broadly as possible.
- Bostrom suggests it should try and *guess* what values the designers have specified for it. It will compute a distribution over many alternative hypotheses, and act accordingly.

Aside: Stuart Russell suggests inverse reinforcement learning.

- Model human agents as having a reward function, and acting so as to maximise reward.
- Then try to infer the reward function from their behaviour.

(ii) Value learning

A problem with any reinforcement learning scheme is wireheading.

(ii) Value learning

A problem with any reinforcement learning scheme is wireheading.

- Rewards are delivered to the agent by a *critic*, that evaluates the current state of the world. (An *actor* then learns to perform actions that maximise present and future rewards.)

(ii) Value learning

A problem with any reinforcement learning scheme is wireheading.

- Rewards are delivered to the agent by a *critic*, that evaluates the current state of the world. (An *actor* then learns to perform actions that maximise present and future rewards.)
- The human designer sets the critic up so that reward states are the ones s/he wants to achieve. (E.g. 'keep the temperature at 20^o' / 'adopt the values you infer from people's behaviour'.)

(ii) Value learning

A problem with any reinforcement learning scheme is wireheading.

- Rewards are delivered to the agent by a *critic*, that evaluates the current state of the world. (An *actor* then learns to perform actions that maximise present and future rewards.)
- The human designer sets the critic up so that reward states are the ones s/he wants to achieve. (E.g. 'keep the temperature at 20^o' / 'adopt the values you infer from people's behaviour'.)
- But now imagine the machine understands its own algorithm, and can modify it...

(ii) Value learning

A problem with any reinforcement learning scheme is wireheading.

- Rewards are delivered to the agent by a *critic*, that evaluates the current state of the world. (An *actor* then learns to perform actions that maximise present and future rewards.)
- The human designer sets the critic up so that reward states are the ones s/he wants to achieve. (E.g. 'keep the temperature at 20^o' / 'adopt the values you infer from people's behaviour'.)
- But now imagine the machine understands its own algorithm, and can modify it...
- The action that maximises reward is no longer the one that pleases the designer, but one that *seizes control of the reward mechanism*.

Philosophy with a deadline

Philosophy with a deadline

‘Before the prospect of an intelligence explosion, we humans are like small children playing with a bomb. Such is the mismatch between the power of our plaything and the immaturity of our conduct.’

Philosophy with a deadline

‘Before the prospect of an intelligence explosion, we humans are like small children playing with a bomb. Such is the mismatch between the power of our plaything and the immaturity of our conduct.’

‘In this situation, any feeling of gee-whiz exhilaration would be out of place. Consternation and fear would be closer to the mark; but the most appropriate attitude may be a bitter determination to be as competent as we can, much as if we were preparing for a difficult exam that will either realise our dreams or obliterate them.’

Philosophy with a deadline

‘Before the prospect of an intelligence explosion, we humans are like small children playing with a bomb. Such is the mismatch between the power of our plaything and the immaturity of our conduct.’

‘In this situation, any feeling of gee-whiz exhilaration would be out of place. Consternation and fear would be closer to the mark; but the most appropriate attitude may be a bitter determination to be as competent as we can, much as if we were preparing for a difficult exam that will either realise our dreams or obliterate them.’

‘This is not a prescription of fanaticism. The intelligence explosion may still be many decades off. (. . .) Yet let us not lose track of what is globally significant. Through the fog of everyday trivialities, we can perceive—if but dimly—the essential task of our age.’

Some strategic suggestions

Some strategic suggestions

Companies / labs / nations are engaged in a *race* to develop AI.

Some strategic suggestions

Companies / labs / nations are engaged in a *race* to develop AI.

- There are rewards for the winner. . .

Some strategic suggestions

Companies / labs / nations are engaged in a *race* to develop AI.

- There are rewards for the winner. . .
- These might encourage players to neglect AI safety research.

Some strategic suggestions

Companies / labs / nations are engaged in a *race* to develop AI.

- There are rewards for the winner. . .
- These might encourage players to neglect AI safety research.

We should encourage *collaboration* between players:

Some strategic suggestions

Companies / labs / nations are engaged in a *race* to develop AI.

- There are rewards for the winner. . .
- These might encourage players to neglect AI safety research.

We should encourage *collaboration* between players:

- to weaken the race dynamic;

Some strategic suggestions

Companies / labs / nations are engaged in a *race* to develop AI.

- There are rewards for the winner. . .
- These might encourage players to neglect AI safety research.

We should encourage *collaboration* between players:

- to weaken the race dynamic;
- to encourage equitable distribution of AI benefits.

Some strategic suggestions

Companies / labs / nations are engaged in a *race* to develop AI.

- There are rewards for the winner. . .
- These might encourage players to neglect AI safety research.

We should encourage *collaboration* between players:

- to weaken the race dynamic;
- to encourage equitable distribution of AI benefits.

The collaboration should relate to AI techniques and AI safety. . .

Some strategic suggestions

Companies / labs / nations are engaged in a *race* to develop AI.

- There are rewards for the winner...
- These might encourage players to neglect AI safety research.

We should encourage *collaboration* between players:

- to weaken the race dynamic;
- to encourage equitable distribution of AI benefits.

The collaboration should relate to AI techniques and AI safety...
But it should *not* result in 'open' AI techniques.

AI researchers should be encouraged to adopt a commitment to AI safety.