

Advice and Perspectives on RL Research Frontiers

Rich Sutton

DeepMind Alberta

University of Alberta

Alberta Machine Intelligence Institute

Reinforcement Learning and Artificial Intelligence Lab



Outline

- Developing your own research thoughts
- A simple trick (completing the square) for doing RL research
- The kind of RL research that I am doing now
- A very different kind of research opportunity: AI & Society

The new RL online course from U Alberta and Coursera



Went live July 25, 2019.
Full specialization
available this Fall

Free!



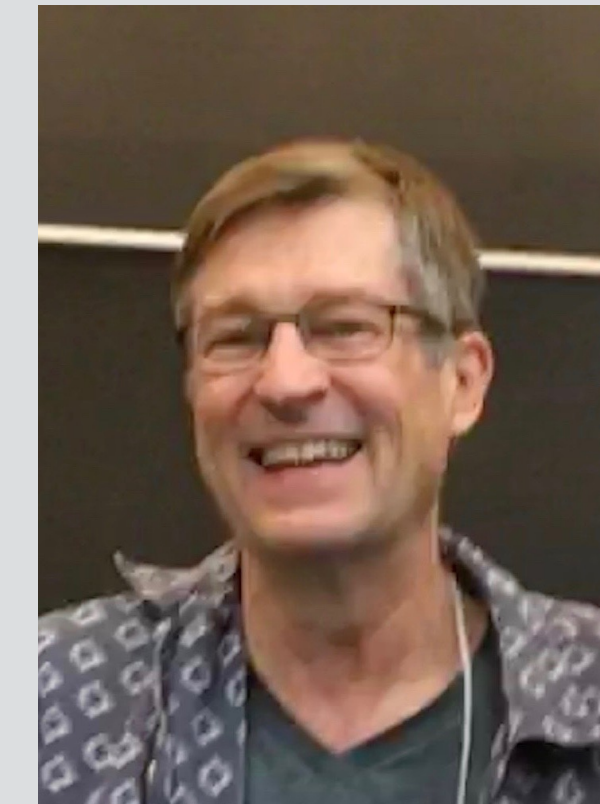
Reinforcement Learning Specialization Instructors Martha White (L) and Adam White

Reinforcement Learning

An Introduction
second edition

Richard S. Sutton and Andrew G. Barto

MIT Press



Andy Barto

Available on the web for free
(<http://richsutton.com>)

There are no authorities in science

- Don't be impressed by what you don't understand
- Don't try to impress others by what they don't understand
- You should be brave and ambitious...
...but also humble and transparent
- Humble before the great task — understanding the mind
 - nature is subtle but not devious
 - it is waiting to be discovered... if we can only see it

Your thoughts are, potentially, of great value

How can you train yourself to
think carefully & productively?

The best way is to write for
yourself

(and discuss with others)

They say it takes 10,000 hours
to become an expert at anything

This could well be true
for thinking about thinking

Are you willing to do the work?

It is not super difficult,
but you do have to show up, day after day



45 years of
my notebooks

A prose poem for your notebook

To write is to begin to think.

To write in a special place,

—a book such as this—

is to honor your thoughts

and to help them build,

one upon the other.

When you get stuck, persist

- In thinking on important questions, you will often reach an apparent dead end, with no where to go
- Here are some techniques for moving forward again:
 - Define your terms
 - Go multiple (What are some of the conceivable answers?)
 - Go meta (What would an answer look like? What properties would it have?)
 - Retreat (to a clearer question that you *can* make progress on)

What is intelligence?

- “Intelligence is the most powerful phenomenon in the universe”
—Ray Kurzweil
- “Intelligence is the computational part of the ability
to achieve goals in the world”
—John McCarthy
- “Intelligence is in the eye of the beholder”

The predictive knowledge hypothesis

“Almost all knowledge of the world can be well thought of as statistics (predictions) about the agent’s future data stream”

Exceptions:

- mathematical knowledge
- knowledge of what to do (policies)
- features
- memories of the past

The most important insight you will ever contribute

- Is probably something that you already know
- Is probably something that is obvious to you
 - *so obvious that you can't see it!*

Sometimes the obvious is the hardest to see.
For example:

- The discovery of gravity, by Isaac Newton
- The discovery that people are animals, evolved from animals, by Charles Darwin
- The discovery of air/vacuum
- The discovery of reinforcement learning by Harry Klopff in the 1970s



Harry Klopff
1941–1997


Are there obvious things that we struggle to see now?

- No animal does supervised learning
- No mind generates images or videos
- Neural networks are not in any meaningful sense “neural”
- People are machines
- The purpose of life is pleasure (and pain)
- The world is much more complex than any mind that tries to understand it
 - therefore, a prior distribution on the world could never be reasonable
- Mind is computational, and computation is increasing exponentially
- Human input doesn't scale; the only scalable methods are search and learning

More advice

- Experience is the data of AI
 - Don't ask the agent to achieve what it can't measure
 - Don't ask the agent to know what it can't verify
- Approximate the solution, not the problem
- Take the agent's point of view
- Set measurable goals for the subparts of an agent
- Work by orthogonal dimensions. Work issue by issue
- Work on ideas, not software

Outline

- Developing your own research thoughts
-  • A simple trick “completing the square” for doing RL research
- The kind of RL research that I am doing now
- A very different kind of research opportunity: AI & Society

The many dimensions of RL

Increasing in difficulty to the right →

- Problem dimensions
 - Prediction — control
 - Bandits — MDPs
 - Discounted — episodic — average reward
 - Fully observable — partially observable
 - Empirical results — convergence theory — rate theory

The frontier??

- Method dimensions
 - Function approximation: tabular — state aggregation — linear — nonlinear
 - Model-free — model-based
 - On-policy — off-policy (Gradient-TD — Emphatic-TD — Tree backup — $Q(\sigma)$ — V-trace)
 - Bootstrapping: Monte Carlo — temporal difference learning
 - Unified treatment by n-step methods — eligibility traces
 - Trace type: Accumulating — replace — dutch — true online
 - Value-based — policy-based
 - State values — action values
 - Batch — online

- Options
- Distributional
- Double and triple methods
- Interest and emphasis

The many dimensions of RL

A good research strategy
is to take two or more dimensions
and “complete the square”

i.e., extend to the right

- Problems dimensions
 - Prediction — control
 - Bandits — MDPs
 - Discounted — episodic — average reward
 - Fully observable — partially observable
 - Empirical results — convergence theory — rate theory
- Method dimensions
 - Function approximation: tabular — state aggregation — linear — nonlinear
 - Model-free — model-based Sutton, Szepesvari, Geramifard, Bowling UAI2008
 - On-policy — off-policy (Gradient-TD — Emphatic-TD — Tree backup — $Q(\sigma)$ — V-trace)
 - Bootstrapping: Monte Carlo — temporal difference learning
 - Unified treatment by n-step methods — eligibility traces
 - Trace type: Accumulating — replace — dutch — true online
 - Value-based — policy-based
 - State values — action values
 - Batch — online
 - Options
 - Distributional
 - Double and triple methods
 - Interest and emphasis

The many dimensions of RL

A good research strategy
is to take two or more dimensions
and “complete the square”

i.e., extend to the right

- Problems dimensions
 - Prediction — control
 - Bandits — MDPs
 - Discounted — episodic — average reward
 - Fully observable — partially observable
 - Empirical results — convergence theory — rate theory
- Method dimensions
 - Function approximation: tabular — state aggregation — linear — nonlinear
 - Model-free — model-based
 - On-policy — off-policy (Gradient-TD — Emphatic-TD — Tree backup — $Q(\sigma)$ — V-trace)
 - Bootstrapping: Monte Carlo — temporal difference learning
 - Unified treatment by n-step methods — eligibility traces
 - Trace type: Accumulating — replace — dutch — true online
 - Value-based — policy-based
 - State values — action values
 - Batch — online
 - Options
 - Distributional
 - Double and triple methods
 - Interest and emphasis

The many dimensions of RL

A good research strategy
is to take two or more dimensions
and “complete the square”

i.e., extend to the right

- Problems dimensions
 - Prediction — control
 - Bandits — MDPs
 - Discounted — episodic — average reward
 - Fully observable — partially observable
 - Empirical results — convergence theory — rate theory
- Method dimensions
 - Function approximation: tabular — state aggregation — linear — nonlinear
 - Model-free — model-based
 - On-policy — off-policy (Gradient-TD — Emphatic-TD — Tree backup — $Q(\sigma)$ — V-trace)
 - Bootstrapping: Monte Carlo — temporal difference learning
 - Unified treatment by n-step methods — eligibility traces
 - Trace type: Accumulating — replace — dutch — true online
 - Value-based — policy-based
 - State values — action values
 - Batch — online
 - Options
 - Distributional
 - Double and triple methods
 - Interest and emphasis


The many dimensions of RL

A good research strategy
is to take two or more dimensions
and “complete the square”

i.e., extend to the right

- Problems dimensions
 - Prediction — control
 - Bandits — MDPs
 - Discounted — episodic — average reward
 - Fully observable — partially observable
 - Empirical results — convergence theory — rate theory
- Method dimensions
 - Function approximation: tabular — state aggregation — linear — nonlinear
 - Model-free — model-based
 - On-policy — off-policy (Gradient-TD — Emphatic-TD — Tree backup — $Q(\sigma)$ — V-trace)
 - Bootstrapping: Monte Carlo — temporal difference learning
 - Unified treatment by n-step methods — eligibility traces
 - Trace type: Accumulating — replace — dutch — true online
 - Value-based — policy-based
 - State values — action values
 - Batch — online
 - Options
 - Distributional
 - Double and triple methods
 - Interest and emphasis

Outline

- Developing your own research thoughts
- A simple trick (completing the square) for doing RL research
-  • The kind of RL research that I am doing now
- A very different kind of research opportunity: AI & Society

The Machine Learning landscape

The old view

- Supervised learning
- Unsupervised learning
- Reinforcement learning

A possible new view

- Prediction learning
- Control learning
- Representation learning
- Integrated agent architectures

The Reinforcement Learning Landscape

In Core RL, we learn

Value functions

Policies

Next, we need to learn

State features

Skills

Models of the world

Subproblems



Animals pursue subproblems that are not the main problem



Babies pursue subproblems that are not the main problem



Babies pursue subproblems that are not the main problem



There is a long history in AI/RL of looking at subproblems that are nominally distinct from the main problem

- Curiosity in RL (Schmidhuber 1991–)
- Multiple learning tasks improves generalization (Caruana 1993-97, Baxter 1997)
- Large numbers of off-policy RL tasks as learning a model of the world (Sutton et al. 1995, 1999, 2011)
- Skills (options) to achieve subgoals (Many 1999–)
- Intrinsic motivation in RL (Barto, Singh, Simsek, Oudeyer, 2005–)
- Auxiliary RL tasks improve generalization (Jaderberg et al. 2014)
- Somewhat settled issues about subproblems:
 - Subproblems are a reward signal and possibly a “terminal” value (subgoal)
 - The solution to a subproblem is an *option* (a policy and a way of terminating)
- Key open questions about subproblems:
 1. What should the subproblems be?
 2. Where do the subproblems come from?
 3. How do the subproblems help the main problem?

How can subproblems help the main problem?

- by shaping the *state representation*
 - feature representations that are good for the subproblem may also be good for the main problem (e.g., distributional RL, auxiliary tasks)
- by shaping *behavior* (making it more coherent or more exploratory)
 - subproblems → options, which are then executed to termination (e.g., the option-critic, termination critic)
- ➔ • by enabling *planning* at a higher level
 - subproblems → options → transition models which are then used in planning (e.g., hierarchical Dyna, Sorg & Singh)
 - planning helps when states change their values (e.g., Airports, Moore and Atkeson, All-goals updating, Kaelbling)

Permanent and transient memories in value function approximation

- One weight vector, the *permanent memory*, is learned in the usual way, say by linear TD(0):

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \alpha \left(R_{t+1} + \gamma \mathbf{w}_t^\top \mathbf{x}_{t+1} - \mathbf{w}_t^\top \mathbf{x}_t \right) \mathbf{x}_t$$

the step size α is chosen to be small, so that the permanent weights converge slowly to the best static approximate value function

- Another weight vector, the transient memory, adds to the permanent weights and learns faster, filling in what the permanent weights miss:

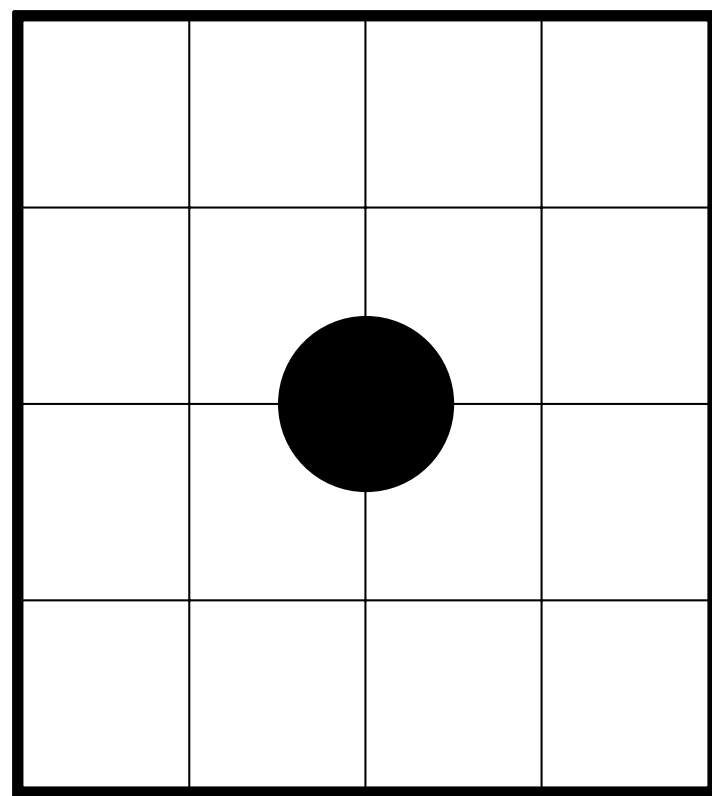
“cascade”

$$\tilde{\mathbf{w}}_{t+1} = \tilde{\mathbf{w}}_t + \tilde{\alpha} \left(R_{t+1} + \gamma (\mathbf{w}_t + \tilde{\mathbf{w}}_t)^\top \mathbf{x}_{t+1} - (\mathbf{w}_t + \tilde{\mathbf{w}}_t)^\top \mathbf{x}_t \right) \mathbf{x}_t$$

where $\tilde{\alpha} > \alpha$. In the long run the transient memory may lose to the permanent memory ($\tilde{\mathbf{w}} \rightarrow \mathbf{0}$), but on tracking problems it can help

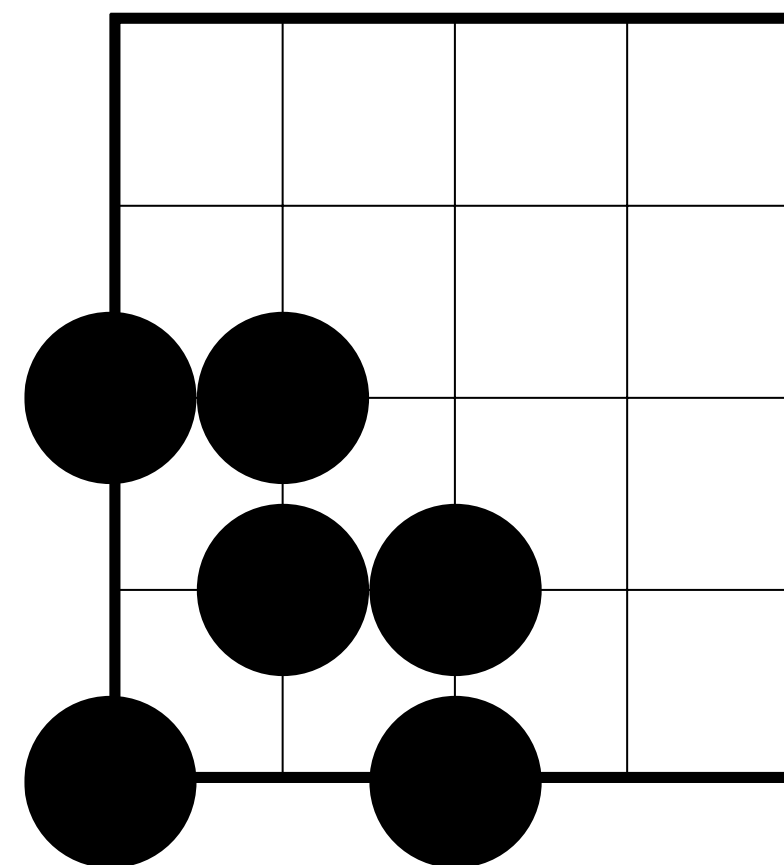
Permanent and transient memories in Go valuation

Local feature
with central
black stone



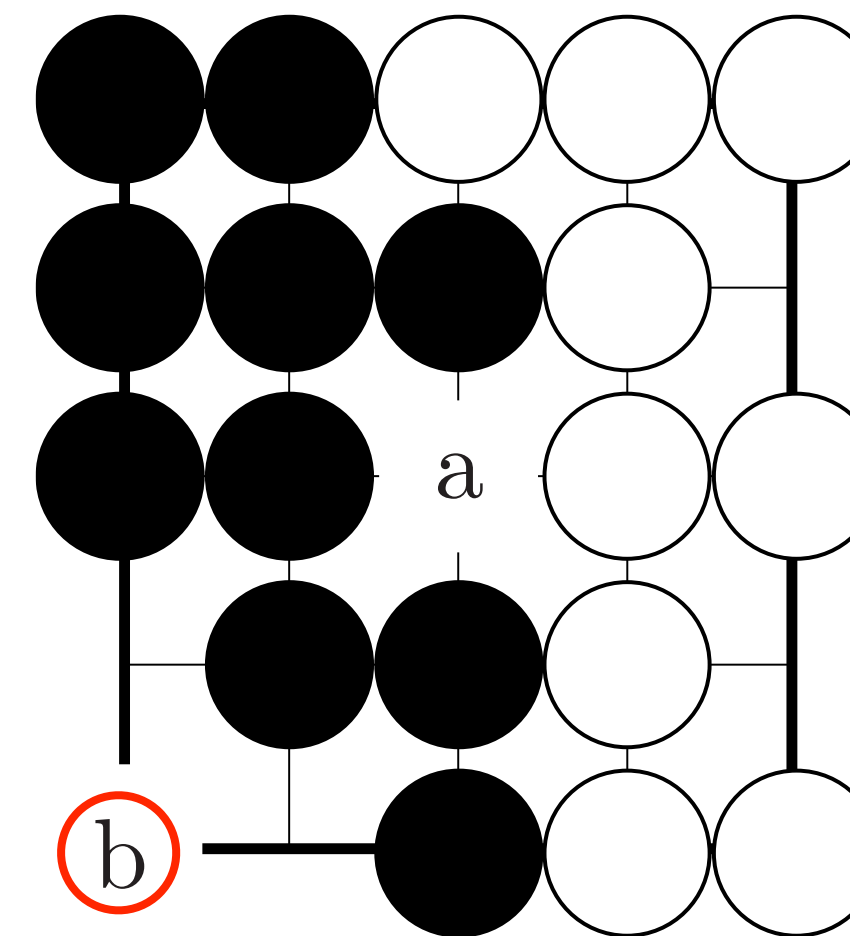
*A strong positive weight
is learned in the
permanent memory
for this feature*

Local feature
with two eyes
in the corner



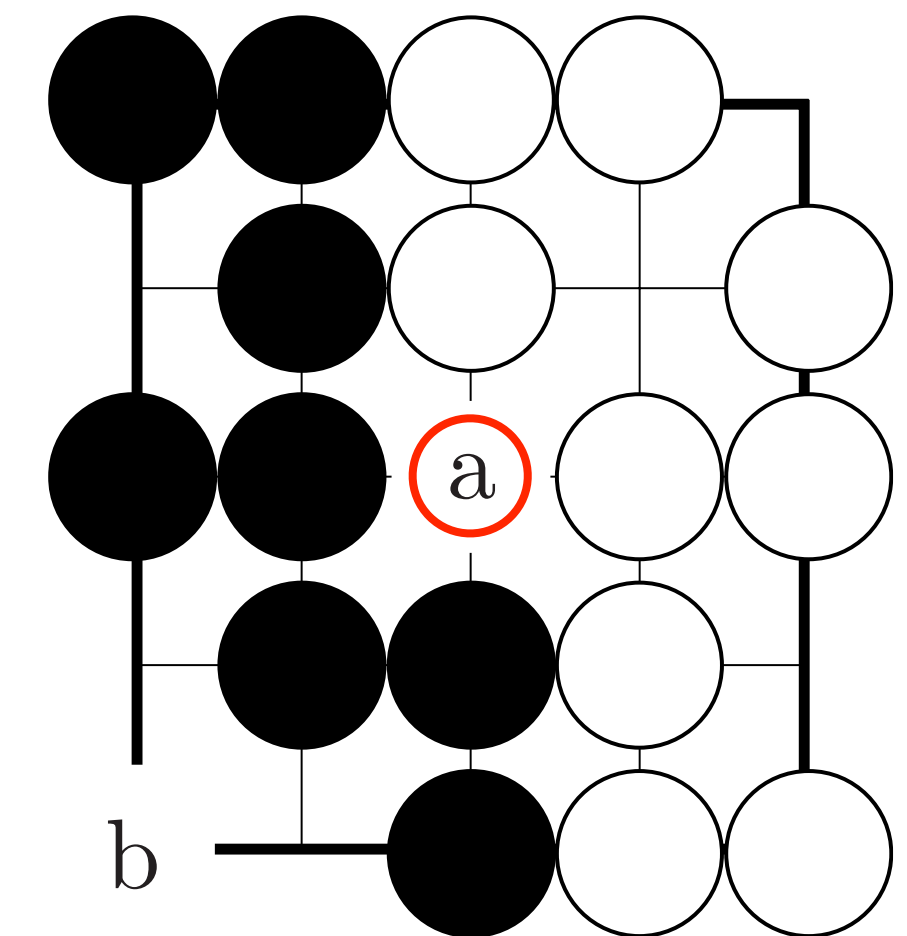
*A modest positive weight
is learned in the
permanent memory
for this feature*

Go positions with two places for black to move.
The permanent memory prefers move a in both



**Move b is winning
here**

The transient
memory learns
this



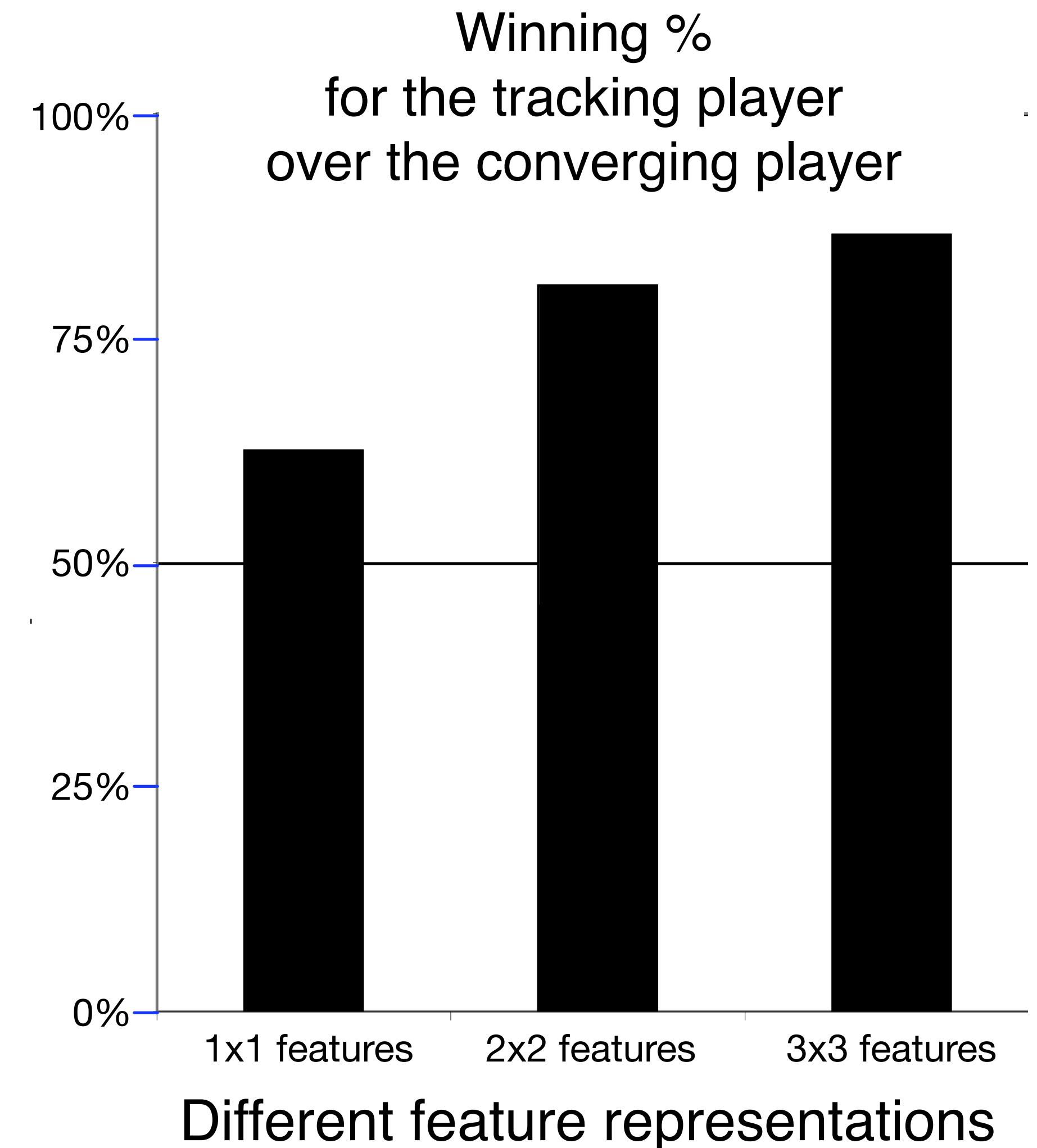
**Move a is winning
here**

The transient
memory does not
interfere

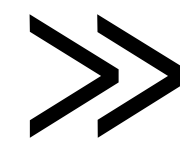
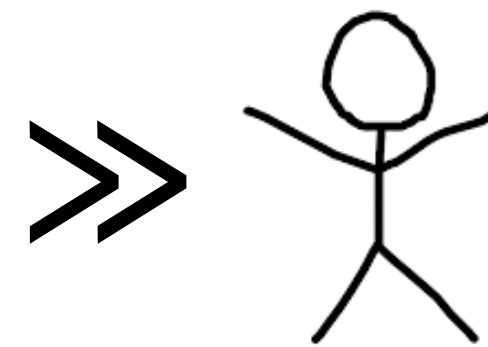
Head-to-head comparison on 5x5 Go

- *Converging player* uses an extensively trained permanent memory to pick moves
- *Tracking player* uses the transient and permanent memories together
- The Tracking player was the clear winner
- The transient memory provides a decisive advantage

It's good to think that values change!



The world is much more complex than the mind



- The mind is too small to contain the exact value function
- There will not be enough weights
- Therefore:
 - We must embrace approximation!
 - The best approximate value function will change even if the world does not

Big world \Rightarrow apparent non-stationarity

\Rightarrow changing *approximate* value function

My answers to the three key open questions about subproblems

1. What should the subproblems be?

Each subproblem should seek to maximize a *single state feature* (then terminate) while respecting the original rewards

Formally, the subproblem for feature i has the same rewards as the usual problem plus, if the option stops at time t , a terminal value of $\mathbf{w}^\top \mathbf{x}_t + x_t^i \cdot \text{Stdev}[\tilde{w}^i]$

2. Where do the subproblems come from?

Subproblems come from state features! There is one subproblem for each feature whose contribution to the value function is highly variable

3. How do the subproblems help on the main problem?

The solution to a subproblem is an option that maximizes its feature; with this, one can act decisively to achieve that feature

And one can learn a transition model of that option, then plan in large abstract steps of feature achievement, as the values of features change

Summary of this approach to integrated RL agents

- A fully capable RL agent must learn larger things—state features, skills, and models—all of which pertain to **subproblems**
- *State-feature achievement*, respecting reward, is a distinctive kind of subproblem
 - that fits well into planning and representation learning
- Because the world is big, we must approximate it;
 - this means it will appear to change, and we will have to track it
 - this is why planning and generalization make sense
- The changes in our approximate value function tell us which features should be the focus of our representations, subproblems, models, and planning

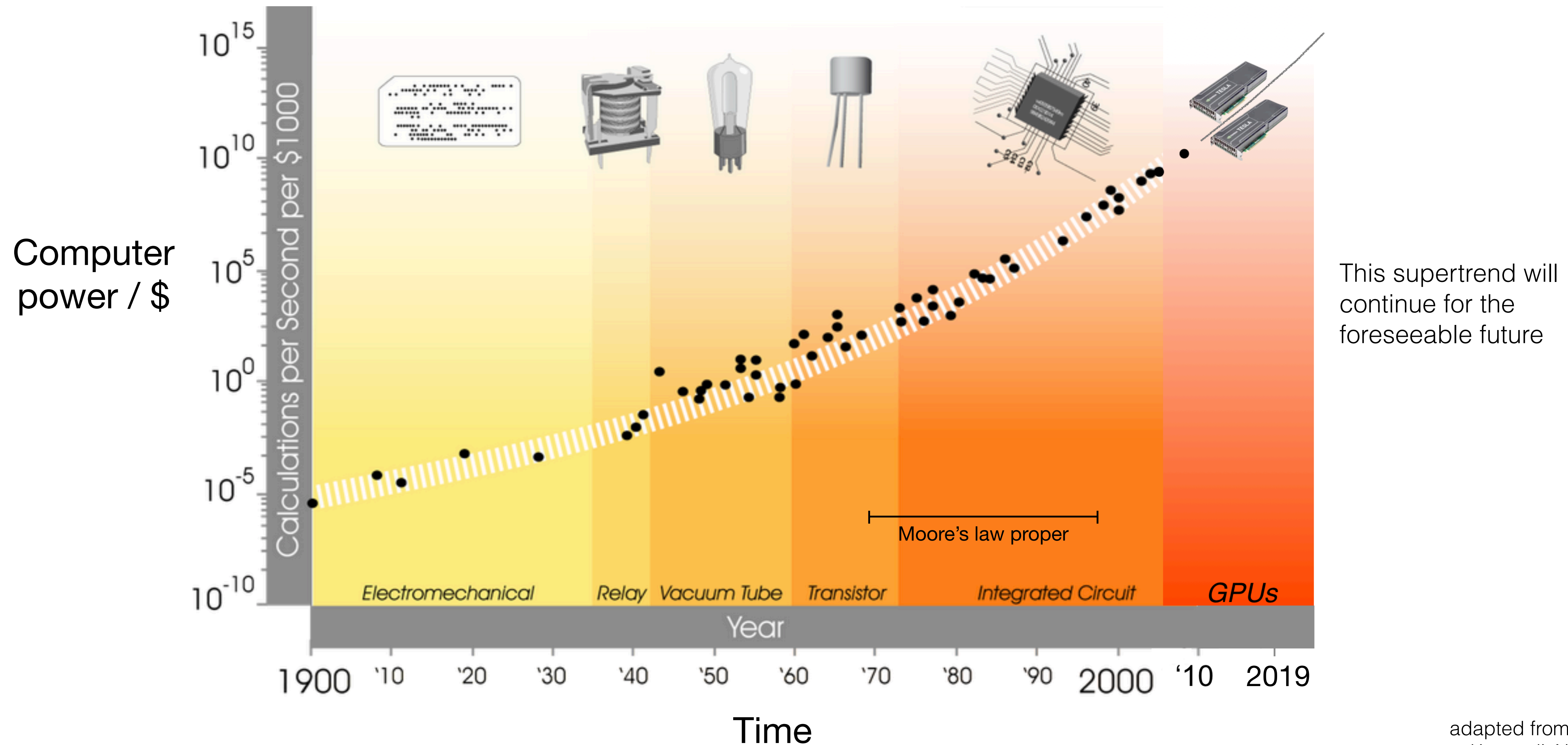
Outline

- Developing your own research thoughts
- A simple trick (completing the square) for doing RL research
- The kind of RL research that I am doing now
- • A very different kind of research opportunity: AI & Society

The coming of AI

- When people finally come to understand the principles of intelligence—what it is and how it works—well enough to design and create beings as intelligent as ourselves
- A fundamental goal for science, engineering, the humanities, ... for all mankind
- It will change the way we work and play, our sense of self, life, and death, the goals we set for ourselves and for our societies
- But it is also of significance beyond our species, beyond history

AI is driven by the supertrend towards ever-cheaper computation (Moore's Law)



AI is the most human-centric of all fields

- It's all about us
 - understanding us, making us, amplifying us
 - not exactly us, but the essential us
 - and making our lives easier, better (that's where the \$ are)
- It is as not techie, alien, artificial, as we make it out to be
- It is us making, or becoming, the next people
- The next step in the evolving, changing, widening river that is ourselves and humankind

Understanding intelligence is surely good, but

- Just understanding intelligence will inevitably lead to ordinary humans falling behind
 - because some people will improve themselves
 - because some people will design improved people
- AI will inevitably lead to new beings and new ways of being that are much more powerful than our current selves

Do unto AIs, as you would have them do unto you

- It is often useful to think of people and AIs as similar
 - both are agents with goals, which may be compatible or conflicting
- So many issues then drop away
 - People should not feel entitlement
 - AIs may not want to be slaves

In the long run...

- AI technology will be part of what disrupts existing social and power structures
 - AIs will force us to re-examine our moral and social foundations
 - Continuing trends that are *1000s of years old*
- AI will bring greater diversities of intelligences, both natural and artificial
 - There will be biases against the new and different.
There will be feelings of entitlement
 - These will be counterproductive and *eventually* fade away
- Will we welcome independent AIs?
- Will we give them a path to joining our society as sovereign persons?

A positive vision of the future

- An open, dynamic, resilient society – peaceful & prosperous
- With a diverse multiplicity of designs, cultures, values, organizations, and sovereign persons of many kinds, both organic and artificial
 - Competing and cooperating
 - With overlapping circles of empathy and support
 - Without feelings of envy or entitlement
- We should care if our design wins, but not insist on it
- The rise of greater foresight in the universe may be one of the few things that is generally good

Good Luck!

and thank you for your attention