Beneficial Smarter-than-human Intelligence: the Challenges and the Path Forward

Benja Fallenstein

Machine Intelligence Research Institute

March 4, 2015

Motivation

- Smarter-than-human intelligence isn't around the corner
  - but it'll (probably) be developed eventually.
- Important to ensure it's **aligned with our interests**
  - But how do we *specify beneficial goals*?
  - How do we make sure system **actually pursues them**?
  - How do we *correct* the system if we get it wrong?
- Want solid **theoretical understanding** of problem & solution
  - Probability theory, decision theory, game theory, statistical learning theory, Bayesian networks, formal verification, ...
  - ...go in the right direction, but *are not enough*.
  - Need for **foundational research**—which can be done today.

Realistic world models

- Contemporary AI systems use simplified models of the world.
  - *e.g.* world state = location of containers and trucks; actions = load container, move truck...
- If you program an agent to pursue a specified goal...
  - ...but that goal wasn't quite right...
  - ...the outcome can be very wrong.
- Idealized description of a physical system *vs.* mathematical model of the entire universe
- If a human smart-aleck can see that your model doesn't match reality, so can a smarter-than-human agent

Solomonoff induction

- Problem: Predict a sequence of bits $x_1, x_2, x_3, \ldots$
  - Given $x_1, \ldots, x_n$, predict $x_{n+1}, x_{n+2}, \ldots$
- Solomonoff induction (roughly):
  - Choose a random program w.p. $\propto 2^{-\text{length}}$
  - Run program to get a sequence of bits
  - Predict by using conditional probabilities
- If the real process generating the sequence is computable
  - then Solomonoff induction predicts well, given enough data
  - But Solomonoff induction itself is uncomputable

Marcus Hutter's AIXI

- Agent interacts with environment
  - In every timestep, agent chooses action a_t
  - Environment responds with observation o_t, reward r_t
  - Problem: Maximize total (time-discounted) reward
- AIXI: Adapt Solomonoff induction. Roughly:
  - Choose random program w.p. $\propto 2^{-\text{length}}$
  - Run program with inputs $a_1, \ldots, a_t$, interpret output as (o_t, r_t)
  - Choose actions maximizing expected discounted reward
- Limitations:
  - Only computable hypotheses
  - AIXI is uncomputable; agent isn't part of the universe
  - No utility function over world states

Reflective oracles

- Is it possible to define an AIXI-like agent which can reason about worlds containing equally powerful agents?
  - Turing machine (TM) can predict other TM by running it...
    - ...but two agents trying to predict each other will loop
  - *Matching pennies:* Two agents choose "heads" or "tails". First agent wins if choose same, second wins if different
    - No deterministic solution
    - Classical game theory solves by *mixed strategies*
- Reflective oracles
  - "Does oracle machine M output 1 w.p. $> p$ when run on this same oracle?"
  - Can answer randomly if probability is exactly p
  - Allows AIXI-like agent to be defined; reproduces Nash equilibria

Vingean reflection

- Can we create a **self-modifying** system. . .
  - . . . that goes through **a billion modifications**. . .
  - . . . *without ever going wrong?*
  - Need *extremely reliable* way for an AI to reason about agents **smarter than itself** — much more reliable than a human!
- Need to use *abstract reasoning*
  - Vinge: Can't know exactly what a smarter successor will do
  - Instead, have *abstract* reasons to think its choices are good
  - Standard decision theory doesn't model this
- Formal logic as a model of abstract reasoning

The "procrastination paradox"

- Agent in a deterministic, known world; discrete timesteps.
- In each timestep, the agent chooses whether to press a button:
  - If pressed in 1^{st} round: Utility $= 1/2$
  - If pressed in 2^{nd} round (and not before): Utility $= 2/3$
  - If pressed in 3^{rd} round (and not before): Utility $= 3/4$
  - ...
  - If never pressed: Utility $= 0$
- (No optimal strategy, but sure can beat 0!)
- The agent is programmed to press the button immediately...
  - ...*unless* it finds a "good argument" that the button will get pressed *later*.

The agent reasons:

- Suppose I don't press the button now.
- Either I press the button in the next step, or I don't.
  - If I *do*, the button gets pressed, good.
  - If I *don't*, I must have found a good argument that the button gets pressed later. So the button gets pressed, good!
  - Either way, the button gets pressed.

So the agent can always find a "good argument" that the button will get pressed later...

- ...and therefore never presses the button!

If we want to have **reliable** *self-referential reasoning, we must understand how to* **avoid this paradox** *(and others like it).*

Logical uncertainty

- Standard probability theory = *environmental* uncertainty.
  - Agents are assumed to be *logically omniscient*.
  - No theoretical understanding of mathematical uncertainty!
- Example: Choose between $O(n^2)$ and $O(n \log n)$ algorithm
- Approach for study:
  - Probability distribution over *complete theories* in some first-order language.
  - *e.g.* complete theories extending Peano Arithmetic (PA)
    - $\rightarrow$ uncertainty about whether PA is consistent
  - Has computable (but very infeasible) analogs

Logical counterfactuals

- Given a world model that makes very accurate predictions. . .
  - . . . and given a utility function exactly modelling our preferences. . .
  - . . . it is still not clear, even in principle, what action an agent should select.
- "Just maximize expected utility. . ."
  - Yes, but how do you compute the expected utility of an action the agent *does not in fact take*?
  - How do you define **what would have happened** in that case?
- Example: Prisoner's Dilemma against isomorphic copy of yourself.
  - Want to cooperate, so that opponent will cooperate.
  - Need counterfactuals that take into account *logical dependencies*.

Conclusions

- Many challenging foundational questions
  - This talk: Realistic world models; Vingean reflection; logical uncertainty; logical counterfactuals
  - Smarter-than-human AI is still in the distant future, but makes sense to begin working on these foundational questions now
  - Hope to build community of researchers in the coming years
- More information:
  - Nick Bostrom: *Superintelligence* (OUP, 2014)
  - `https://intelligence.org/technical-agenda/`

Thank you for your attention!