A game theoretic look at alignment

Dean Foster, Dhruv Madeka, Sham Kakade

September 11, 2023

Game Theory: Interacting Decision Makers

Game theory is about interactive decision making:

- It has very little to do with Chess and checkers!
- But lots to do with:
  - evolution
  - knowledge
  - manipulation
  - deception
  - reputation
  - trust
  - reputation
  - communication
- All ripe areas for modeling alignment

Game Theory: Interacting Decision Makers

Game theory is about interactive decision making:

- It has very little to do with Chess and checkers!
- But lots to do with:
  - evolution
  - knowledge
  - manipulation
  - deception
  - reputation
  - trust
  - reputation
  - communication
- All ripe areas for modeling alignment

I'll take questions until slide 21!

Connection to security

Many similarities with security:

- Randomization:
  - games: necessary for games to protect private knowledge
  - CS: necessary for interactive proofs and zero knowledge proofs
- Chains of reputation:
  - games: Useful for identifying bad actors
  - CS: "web of trust"
- Openness is better:
  - games: mechanism design
  - CS: security through obscurity isn't secure

Trust

- Consider an “executive” of a company
  - The compay trusts the executive with the power to buy start-ups
  - But the company gives them zero training

Trust

- Consider an "executive" of a company
  - The compay trusts the executive with the power to buy start-ups
  - But the company gives them zero training
- The company doesn't trust the executive to log into their email!

Trust

- Consider an “executive” of a company
  - The compay trusts the executive with the power to buy start-ups
  - But the company gives them zero training
- The company doesn’t trust the executive to log into their email!
  - THey need two factor authentication to log in
  - Two factors aren’t necessary to buy a startup!

Humans trust too much

A few years ago I got scammed on the street by being told a sob story.

Humans trust too much

A few years ago I got scammed on the street by being told a sob story.

- Like many humans, I trust other people too much
- After it happened, I decided I was comfortable being a schmuck since the alternative was to trust less
- So knowing when human's will stupidly "trust" is an issue for alignment

Humans trust too much

A few years ago I got scammed on the street by being told a sob story.

- Like many humans, I trust other people too much
- After it happened, I decided I was comfortable being a schmuck since the alternative was to trust less
- So knowing when human's will stupidly "trust" is an issue for alignment
- (If Chimps ruled the world, we wouldn't have to worry about alignment–they trust no-one!)

The company's policy makes sense

- The company knows the executive is susceptible to spear phishing
  - So they lock that door twice!
- They know the executive won't trust a valuation of a start up as being a "good deal"
  - So they don't even lock that door once

TRUST

Connection to security

Many similarities with security:

- Randomization:
  - games: necessary for games to protect private knowledge
  - CS: necessary for interactive proofs and zero knowledge proofs
- Chains of reputation:
  - games: Useful for identifying bad actors
  - CS: "web of trust"
- Openness is better:
  - games: mechanism design
  - CS: security through obscurity isn't secure

Trust

- Consider an "executive" of a company
  - The compay trusts the executive with the power to buy start-ups
  - But the company gives them zero training

Humans trust too much

A few years ago I got scammed on the street by being told a sob story.

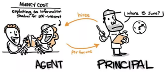
The company's policy makes sense

- The company knows the executive is susceptible to spear phishing
  - So they lock that door twice!
- They know the executive won't trust a valuation of a start up as being a "good deal"
  - So they don't even lock that door once

Information vs. computation

- In game theory, all true facts are common knowledge
- We will model computation as information

The Principal / Agent problem



Evolution

Oldest example of Principal / agent:

- Flowers and bees!
- Flowers "pay" bees to pollinate for them
- Flower is principal
- Bee is agent
- The deal:
  - Payment in nectar
  - Paid half in advance and half afterwards
  - Variable payment based on number of bees in the market place
  - Successful arrangement for 100 million years
- Note: Bees are *much* smarter than flowers

Farming: Share cropping

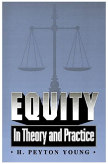
- Principal: Land owner
- Agent: Farmer
- The deal:

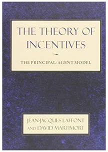
:ffort vs

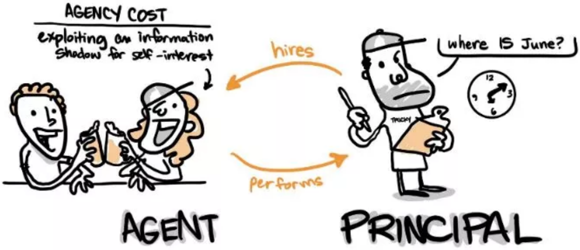
Theory: Agents have knowledge

- Agents know more than principals
  - Necessary for game theory model
  - Otherwise, principal can simply pay "piece work"
- We will be modeling super-AIs as more knowledgeable
  - knowledge in game theory is sigma-fields, observations from the world, knowledge of ones personal utility function, etc
  - None of these apply to an AI
  - But they are better at computation
  - Which looks a lot like information
  - We will take it as being the same

Books:







AGENCY COST
exploiting an information shadow for self-interest
hires
where IS June?
performs
AGENT
PRINCIPAL

OUR GOAL IS TO WRITE BUG-FREE SOFTWARE. I'LL PAY A TEN-DOLLAR BONUS FOR EVERY BUG YOU FIND AND FIX.
S. Adams E-mail: SCOTTADAMS@AOL.COM
YAHOO!
WE'RE RICH
YES!!! YES!!! YES!!!
1/13 © 1995 United Feature Syndicate, Inc.(NYC)
I HOPE THIS DRIVES THE RIGHT BEHAVIOR.
I'M GONNA WRITE ME A NEW MINIVAN THIS AFTER-NOON!

Evolution

Oldest example of Principal / agent:

- Flowers and bees!
- Flowers "pay" bees to pollinate for them
- Flower is principal
- Bee is agent
- The deal:
  - Payment in nectar
  - Paid half in advance and half afterwards
  - Variable payment based on number of bees in the market place
  - Successful arrangement for 100 million years
- Note: Bees are *much* smarter than flowers

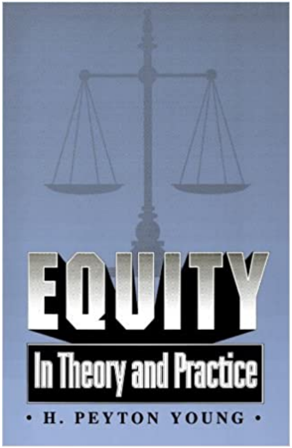
Farming: Share cropping

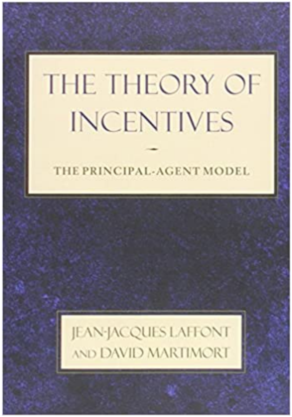
- Principal: Land owner
- Agent: Farmer
- The deal:
  - Farmers give half of the proceeds to owner
  - Owner doesn't know how much productivity is due to effort vs luck
  - 50 / 50 split is common, but other splits are possible
- Note: Owners don't have to know farming

Theory: Agents have knowledge

- Agents know more than principals
  - Necessary for game theory model
  - Otherwise, principal can simply pay “piece work”
- We will be modeling super-AIs as more knowledgeable
  - knowledge in game theory is sigma-fields, observations from the world, knowledge of ones personal utility function, etc
  - None of these apply to an AI
  - But they are better at computation
  - Which looks a lot like information
  - We will take it as being the same

Books:


EQUITY
In Theory and Practice
• H. PEYTON YOUNG •


THE THEORY OF
INCENTIVES
THE PRINCIPAL-AGENT MODEL
JEAN-JACQUES LAFFONT
AND DAVID MARTIMORT

Game Theory Questions?

TRUST

Connection to security

Many similarities with security:
- Randomization:
  - games: necessary for games to protect private knowledge
  - CS: necessary for interactive proofs and zero knowledge proofs
- Chains of reputation:
  - games: Useful for identifying bad actors
  - CS: "web of trust"
- Openness is better:
  - games: mechanism design
  - CS: security through obscurity isn't secure

Trust

- Consider an "executive" of a company
  - The compay trusts the executive with the power to buy start-ups
  - But the company gives them zero training

Humans trust too much

A few years ago I got scammed on the street by being told a sob story.

The company's policy makes sense

- The company knows the executive is susceptible to spear phishing
  - So they lock that door twice!
- They know the executive won't trust a valuation of a start up as being a "good deal"
  - So they don't even lock that door once

Information vs. computation

- In game theory, all true facts are common knowledge
- We will model computation as information

The Principal / Agent problem

Evolution

Oldest example of Principal / agent:
- Flowers and bees!
- Flowers "pay" bees to pollinate for them
- Flower is principal
- Bee is agent
- The deal:
  - Payment in nectar
  - Paid half in advance and half afterwards
  - Variable payment based on number of bees in the market place
  - Successful arrangement for 100 million years
- Note: Bees are much smarter than flowers

Farming: Share cropping

- Principal: Land owner
- Agent: Farmer

Theory: Agents have knowledge

- Agents know more than principals
  - Necessary for game theory model
  - Otherwise, principal can simply pay "piece work"
- We will be modeling super-AIs as more knowledgeable
  - knowledge in game theory is sigma-fields, observations from the world, knowledge of ones personal utility function, etc
  - None of these apply to an AI
  - But they are better at computation
  - Which looks a lot like information
  - We will take it as being the same

Books:

That was slide 21!

Using game theory, I'll argue for the following policy suggestions:

That was slide 21!

Using game theory, I'll argue for the following policy suggestions:

Policy suggestions:

- Launch early
- Launch many
- Private AIs are unregulated (e.g. tutors / advobots)
- Public AIs:
  - log all their statements (block-chain AI?)
  - AIs are tiered / cross checked

Launching early: Trust

Humans need to learn lack of trust:

- 1890's yellow journalism (modern tabloids)
- 1950's chain letters and mail fraud
- 1990's email chain letters (lead to snoops)
- 2010's Facebook for "real news"

Launching early: Trust

Humans need to learn lack of trust:

- 1890's yellow journalism (modern tabloids)
- 1950's chain letters and mail fraud
- 1990's email chain letters (lead to snoops)
- 2010's Facebook for "real news"
- 2020's AI

So launching earlier will allow humans to get used to them

Pox parties

- We need to throw chicken pox parties!
  - These were common when I was a kid
  - We'd go to a sick child's house and hopefully get chicken pox
  - Hopefully no one under 30 has a clue what I'm talking about
  - (Vaccine came out in 1995)

Pox parties

- We need to throw chicken pox parties!
  - These were common when I was a kid
  - We'd go to a sick child's house and hopefully get chicken pox
  - Hopefully no one under 30 has a clue what I'm talking about
  - (Vaccine came out in 1995)
- We have no vaccine against evil AIs
- We need to get inoculated by exposure to real AIs
- Hopefully we can build up immunity as we progress from GPT4, 5, 6, ...

Launching early: Learning

- Real game theorist solve games backwards

Launching early: Learning

- Real game theorist solve games backwards
- I'm not a real game theorist!
  - Neither are most animals or humans
  - We learn from experience
  - Use that for future interactions

Launching early: Learning

- Real game theorist solve games backwards
- I'm not a real game theorist!
  - Neither are most animals or humans
  - We learn from experience
  - Use that for future interactions
- But, won't super smart AIs learn faster than humans if we have repeated interactions?

Aside: Repeated games

- If a FSA(n) plays a FSA(2^n) it loses.[1]

[1] Actually, maybe it is FSA(2^{2^n}) but who's counting?

Aside: Repeated games

- If a FSA(n) plays a FSA(2^n) it loses.[1]
- But, if a a FSA(O(1)) is allowed to toss a coin, then it plays well against an arbitrarily smart adversary.

[1] Actually, maybe it is FSA(2^{2^n}) but who's counting?

Aside: Repeated games

- If a FSA(n) plays a FSA(2^n) it loses.[1]
- But, if a a FSA(O(1)) is allowed to toss a coin, then it plays well against an arbitrarily smart adversary.
- This is true, even if the stupid FSA has to learn the correct strategy to play. (F. and Vohra 1998, F. and Kakade 2008)

[1] Actually, maybe it is FSA(2^{2^n}) but who's counting?

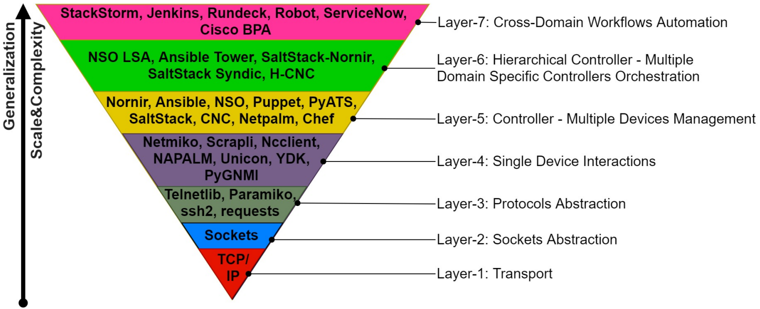
Launch early

Launching early is a win because we:

- learn appropriate trust
- builds immunity
- learning doesn't favor the more intelligent

Principal / Agent

Network Automation Abstraction Layers Taxonomy



GPT4 as middle manager

- GPT4 can understand GPT5
  - Model GPT4 as having more information than we humans have
  - Use σ-fields
- Humans can understand GPT4
  - align GPT4's goals with human goals
  - Let GPT4 figure out how to align GPT5
- No trust is needed!

Mathematics

- Human's σ-field is $\mathcal{F}_0$.
- GPT4's σ-field is $\mathcal{F}_4$.
- GPT5's σ-field is $\mathcal{F}_5$.
- GPT5 knows more than GPT4 which knows more than the human:

$$\mathcal{F}_0 \subset \mathcal{F}_4 \subset \mathcal{F}_5$$

- $A_0 \in \mathcal{F}_0$.
- $A_4 \in \mathcal{F}_4$.
- $A_5 \in \mathcal{F}_5$.
- $E(U_0(\vec{A})|\mathcal{F}_0) \in \mathcal{F}_0$.
- Exotic Assumptions:
  - $E(U_4(\vec{A})|\mathcal{F}_4) \in \mathcal{F}_0$.
  - $E(U_5(\vec{A})|\mathcal{F}_5) \in \mathcal{F}_4$.

Theorem

In this middle management principal agent model, the human's goals are aligned with GPT5's goals.

Launching many: So they can control each other

Many player games are easy

- Multiplayer games don't require as much strategic thinking
- An "economy of agents" is easier than a single agent

Many player games are easy

- Multiplayer games don't require as much strategic thinking
- An "economy of agents" is easier than a single agent
- So, having many AIs is better than having a few
- Again: launch many!

Launch many

Launching many is a win because:

- middle management / indirection
- economy requires less strategy than game theory

Pseudo randomization

- Stackelberg equilibrium
- Example: Amazon vs FBA sellers
  - Each seller acts like a "random draw"
  - Amazon has to have a single policy for all sellers
- One AI against many people
  - pre-commit to what it is saying
  - Force it to tell a consistent story
  - Logging its statements
- TCS version: PCP

Putting this together

- Launch early:
  - trust / reputation
  - builds immunity
  - learning
- Launch many:
  - economies are simpler than games (MIPs)
  - middle management
- Personalized private copies:
  - force privacy to avoid collusion
- large LLMs log their statements:
  - Stackelberg equilibrium (PCP)

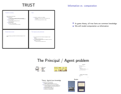
Final thoughts

- Game theory is useful model of human / AI interactions
  - Evolution has been solving these problems for billions of years
  - Humans have been solving them for millions of years
  - Legal codes have been solving them for 1000s of years
  - We can use this accumulated knowledge for alignment

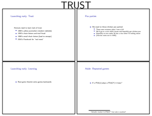
THANKS!

THANKS!

Game Theory Questions?



TRUST



Many player games are easy

- Multiplayer games don't require as much strategic thinking
- An "economy of agents" is easier than a single agent

TRUST

Launching early: Trust

Humans need to learn lack of trust:

- 1890's yellow journalism (modern tabloids)
- 1950's chain letters and mail fraud
- 1990's email chain letters (lead to snoops)
- 2010's Facebook for "real news"

Pox parties

- We need to throw chicken pox parties!
  - These were common when I was a kid
  - We'd go to a sick child's house and hopefully get chicken pox
  - Hopefully no one under 30 has a clue what I'm talking about
  - (Vaccine came out in 1995)

Launching early: Learning

- Real game theorist solve games backwards

Aside: Repeated games

- If a FSA(n) plays a FSA(2^n) it loses.[1]

[1] Actually, maybe it is FSA(2^{2^n}) but who's counting?

Mathematics

- Human's σ-field is $\mathcal{F}_0$.
- GPT4's σ-field is $\mathcal{F}_4$.
- GPT5's σ-field is $\mathcal{F}_5$.
- GPT5 knows more than GPT4 which knows more than the human:

$$\mathcal{F}_0 \subset \mathcal{F}_4 \subset \mathcal{F}_5$$

- $A_0 \in \mathcal{F}_0$.
- $A_4 \in \mathcal{F}_4$.
- $A_5 \in \mathcal{F}_5$.
- $E(U_0(\vec{A})|\mathcal{F}_0) \in \mathcal{F}_0$.
- Exotic Assumptions:
  - $E(U_4(\vec{A})|\mathcal{F}_4) \in \mathcal{F}_0$.
  - $E(U_5(\vec{A})|\mathcal{F}_5) \in \mathcal{F}_4$.

Theorem

In this middle management principal agent model, the human's goals are aligned with GPT5's goals.