Vectoring in Research

CS 197 | Stanford University | Brando Miranda
cs197.stanford.edu

Slides adapted from Kanishk Gandhi & Michael Bernstein

Vectoring in Research

CS 197 | Stanford University | Brando Miranda
cs197.stanford.edu

Slides adapted from Kanishk Gandhi & Michael Bernstein

Administrivia

You don’t have any assignment for the next few weeks, you are now set to dive deeper into research!

What problem are we solving?

“But how do we start?”

“I’m feeling so lost.”

“I thought of an important reason that this won’t work.”

“It’s not working yet. I’m not sure that we’re making progress.”

Today's big idea: vectoring

What is vectoring?

How do we vector effectively?

What goes wrong if we don't vector?

See Brando's prompt for LLMs to discuss vectors in your research: https://gist.github.com/brando90/00e5e3c66f5349a0c3cbc62ef2501904

Feel free to leave a comment and try to improve it!

Michael’s theory of Researcher success

To be a successful researcher, you need to master two skills that operate in a *tight loop* with one another.

Vectoring: identifying the biggest dimension of risk in your project right now (often assumption/wrt to main objective/H)

oday!

Velocity: rapid reduction of risk in the chosen dimension (you want to learn ASAP – you don’t want to “build your life on a lie”! e.g., build it manually vs whole infra)

not today!

What Is Vectoring?

What we think research is

Credit: Stefan Savage, https://dl.acm.org/doi/abs/10.1145/3517745.3570969

What we think research is

Credit: Stefan Savage, https://dl.acm.org/doi/abs/10.1145/3517745.3570969

What research is not

1. Figure out what to do.
2. Do it.
3. Publish.

What research is

Research is an iterative process of exploration, not a linear path from idea to result [Gowers 2000]

Problematic points of view

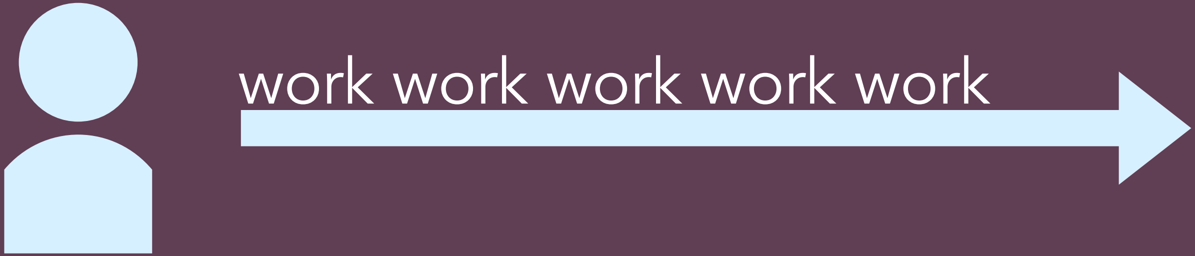
"OK, we have a good idea. Let's build it / model it / prove it / get training data."

"I spent some time thinking about this and hacking on it, and it's not going to work: it has a fatal flaw."

Problem: Treating your research goal as a project spec and executing it

Idea as project spec

Taking a concept and trying to realize it in parallel across all decisions, assumptions, and goals



work work work work work

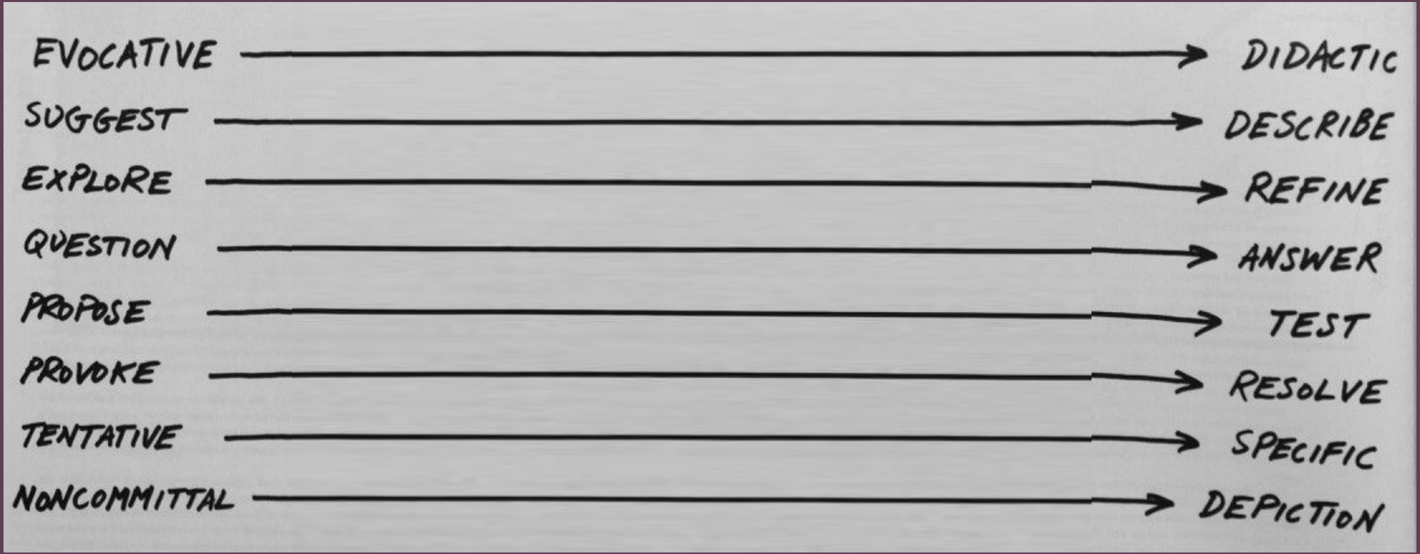
Result

Concept

Idea as project spec

What you should have done

What you did



EVOCATIVE → DIDACTIC

SUGGEST → DESCRIBE

EXPLORE → REFINE

QUESTION → ANSWER

PROPOSE → TEST

PROVOKE → RESOLVE

TENTATIVE → SPECIFIC

NONCOMMITTAL → DEPICTION

[Buxton 2007]

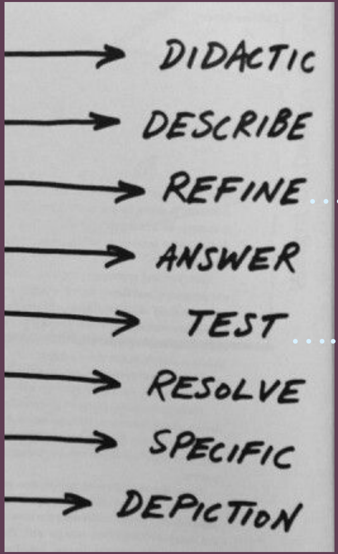
This is all other points
of a research project & process
Iteratively derisk & explore/learn

This is the endpoint
of a research project
(end goal to communicate)

Problematic points of view

"OK, we have a good idea. Let's build it / model it / prove it / get training data."

"I spent some time thinking about this and hacking on it, and it's not going to work: it has a fatal flaw."



……before knowing what to refine!

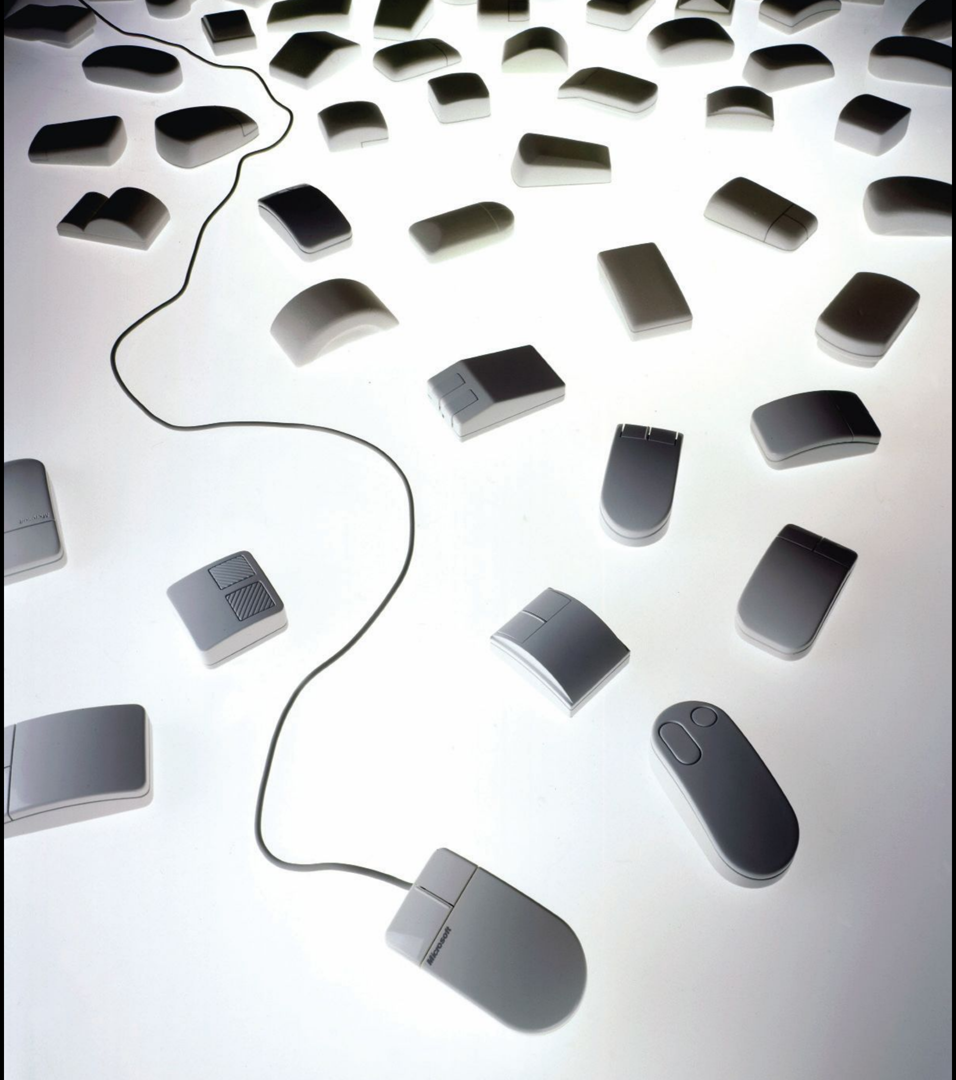
……before identifying if that test or flaw is the right one to focus on!

Pick a vector - dimension risk

It may feel like we get stuck, unable to solve the problem because we haven’t figured out everything (perfection!) else about it.
There are too many open questions, and too many possible directions.
The more dimensions there are, the harder gradient descent becomes.

Instead of trying to do everything at once (project spec), pick *one* dimension of uncertainty — one vector — and focus on reducing its risk and uncertainty.

Scope your vector to be something you can reduce uncertainty on in 1–2 weeks


Microsoft

Example vectors

Piloting: will this technique work at all? To answer this, we implement a *basic* version of the technique and mock in the data and other test harness elements.

Engineering: will this technique work with a realistic workload? To answer this, we need to engineer a test harness.

Proving: does this limit that I suspect does? To answer this, we start by writing a proof for a simpler case.

Design: what might this interaction look like to an end user? To answer this, we create a low-fi prototype.

Implications

The vectors under consideration will each imply building different parts of your system.

Rather than building them all at once, when you might have to change things later, vectoring instead implies that you start by reducing uncertainty in the most important dimension first — your “inner loop” — and then building out from there.

Vectoring algorithm

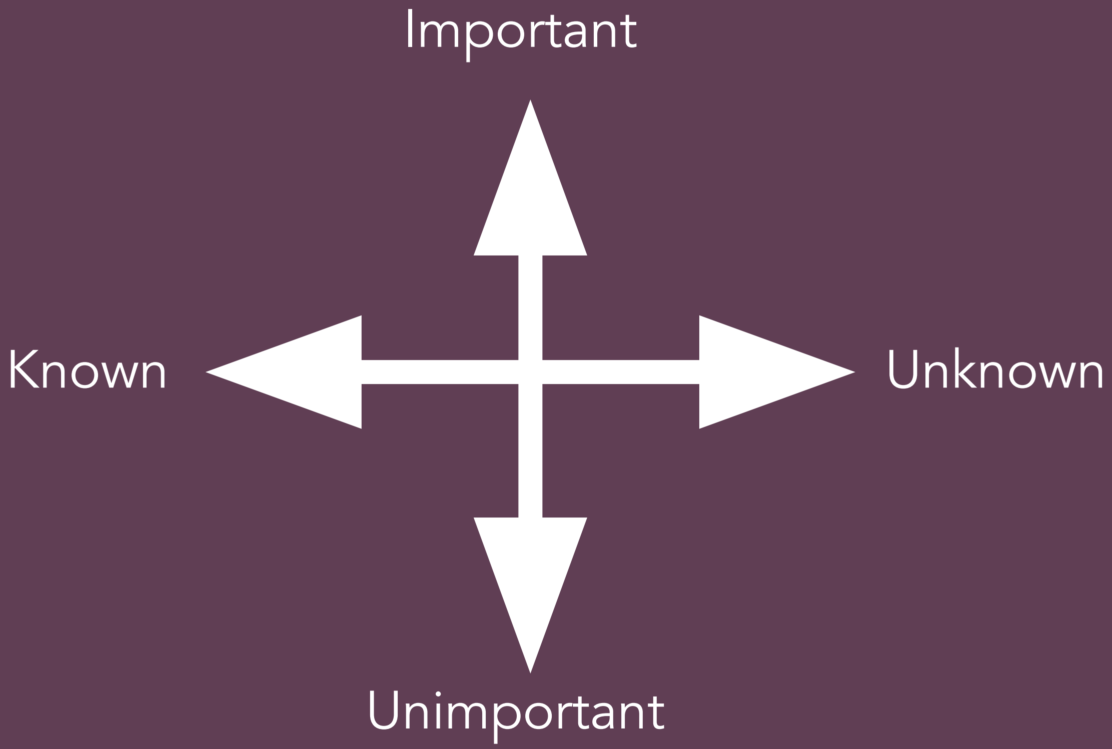
1. Generate questions
Untested hunches, risky decisions,
high-level directions

2. Rank your questions
Which is most critical?

3. Pick one and answer it rapidly
Answer only the most critical question
(This is where velocity comes into play)

Assumption mapping

Assumption mapping is a strategy for articulating questions and ranking them.



Let’s Try It

Emergence in LLMs?

Assumption: Everyone thinks emergent capabilities (sharp unpredictable jumps in performance) of LLMs is a fundamental property of scaling AI models

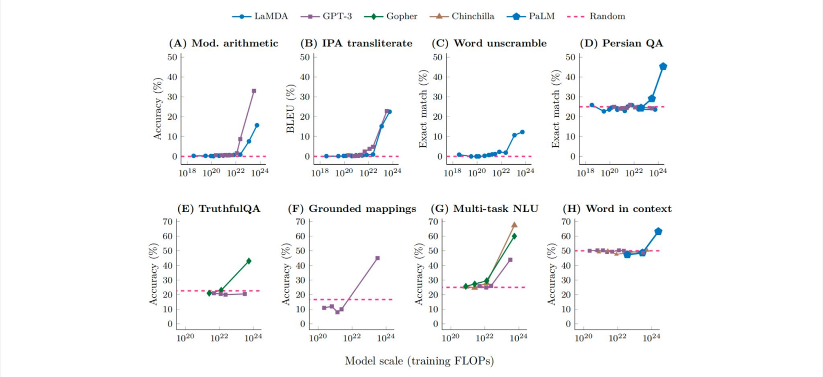
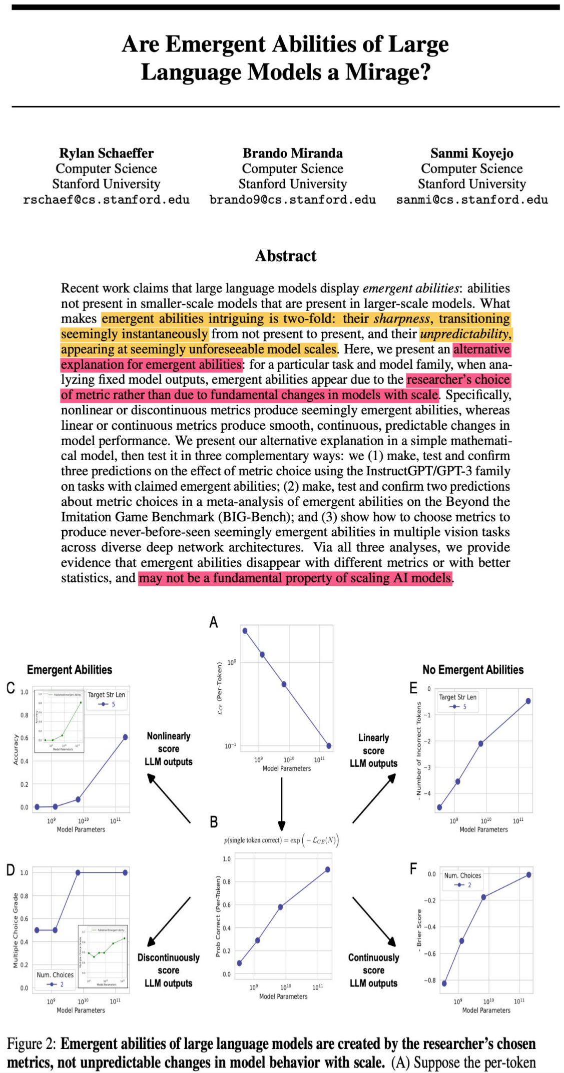


Figure 1: **Emergent abilities of large language models**. Model families display *sharp* and *unpredictable* increases in performance at specific tasks as scale increases. Source: Fig. 2 from [38].

Hypothesis: authors had a hunch it was mainly due to other factors



Are Emergent Abilities of Large Language Models a Mirage?

Rylan Schaeffer
Computer Science
Stanford University
rschaef@cs.stanford.edu

Brando Miranda
Computer Science
Stanford University
brando9@cs.stanford.edu

Sanmi Koyejo
Computer Science
Stanford University
sanmi@cs.stanford.edu

Abstract

Recent work claims that large language models display *emergent abilities*: abilities not present in smaller-scale models that are present in larger-scale models. What makes emergent abilities intriguing is two-fold: their *sharpness*, transitioning seemingly instantaneously from not present to present, and their *unpredictability*, appearing at seemingly unforeseeable model scales. Here, we present an alternative explanation for emergent abilities: for a particular task and model family, when analyzing fixed model outputs, emergent abilities appear due to the researcher's choice of metric rather than due to fundamental changes in models with scale. Specifically, nonlinear or discontinuous metrics produce seemingly emergent abilities, whereas linear or continuous metrics produce smooth, continuous, predictable changes in model performance. We present our alternative explanation in a simple mathematical model, then test it in three complementary ways: we (1) make, test and confirm three predictions on the effect of metric choice using the InstructGPT/GPT-3 family on tasks with claimed emergent abilities; (2) make, test and confirm two predictions about metric choices in a meta-analysis of emergent abilities on the Beyond the Imitation Game Benchmark (BIG-Bench); and (3) show how to choose metrics to produce never-before-seen seemingly emergent abilities in multiple vision tasks across diverse deep network architectures. Via all three analyses, we provide evidence that emergent abilities disappear with different metrics or with better statistics, and may not be a fundamental property of scaling AI models.

A
Emergent Abilities
No Emergent Abilities
C
E
Target Str Len
5
Accuracy
$\mathcal{L}_{CE}$ (Per-Token)
- Number of Incorrect Tokens
Model Parameters
Nonlinearly score LLM outputs
Linearly score LLM outputs
B
$p(\text{single token correct}) = \exp\left(-\mathcal{L}_{CE}(N)\right)$
D
F
Multiple Choice Grade
Prob Correct (Per-Token)
- Brier Score
Num. Choices
2
Discontinuously score LLM outputs
Continuously score LLM outputs
Published Emergent Ability

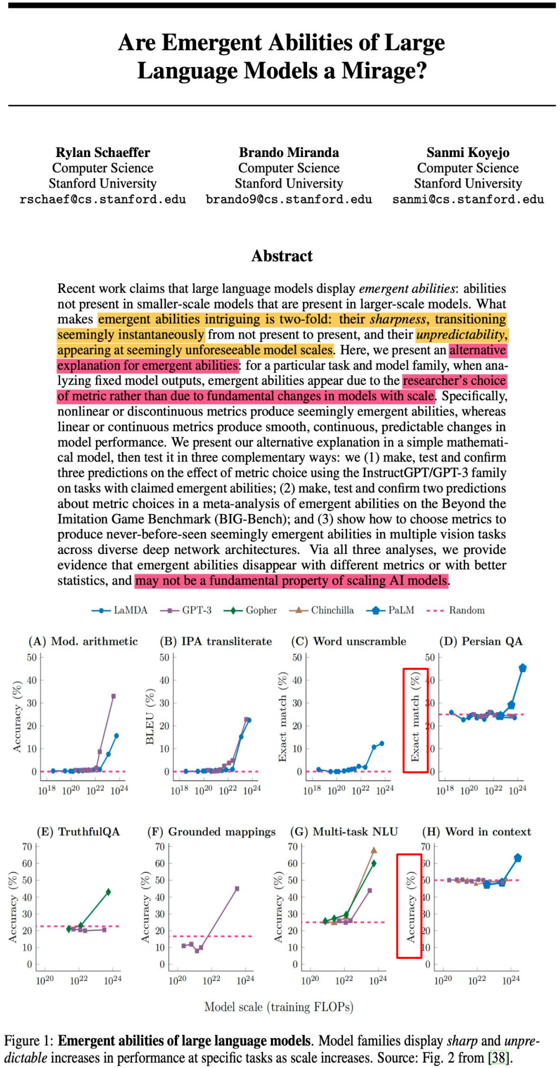
Figure 2: **Emergent abilities of large language models are created by the researcher's chosen metrics, not unpredictable changes in model behavior with scale.** (A) Suppose the per-token

Emergence in LLMs?

Hypothesis: Emergent Capabilities (unpredictable jumps) were due to different factors than fundamental properties of scaling AI models

Possible Vectors:

1. Is it due to model scoring metric?
2. Is it due to low number of sample in test set?
3. Is it due to sparse sampling of model scales?



Are Emergent Abilities of Large Language Models a Mirage?

Rylan Schaeffer
Computer Science
Stanford University
rschaef@cs.stanford.edu

Brando Miranda
Computer Science
Stanford University
brando9@cs.stanford.edu

Sanmi Koyejo
Computer Science
Stanford University
sanmi@cs.stanford.edu

Abstract

Recent work claims that large language models display *emergent abilities*: abilities not present in smaller-scale models that are present in larger-scale models. What makes emergent abilities intriguing is two-fold: their *sharpness*, transitioning seemingly instantaneously from not present to present, and their *unpredictability*, appearing at seemingly unforeseeable model scales. Here, we present an alternative explanation for emergent abilities: for a particular task and model family, when analyzing fixed model outputs, emergent abilities appear due to the researcher's choice of metric rather than due to fundamental changes in models with scale. Specifically, nonlinear or discontinuous metrics produce seemingly emergent abilities, whereas linear or continuous metrics produce smooth, continuous, predictable changes in model performance. We present our alternative explanation in a simple mathematical model, then test it in three complementary ways: we (1) make, test and confirm three predictions on the effect of metric choice using the InstructGPT/GPT-3 family on tasks with claimed emergent abilities; (2) make, test and confirm two predictions about metric choices in a meta-analysis of emergent abilities on the Beyond the Imitation Game Benchmark (BIG-Bench); and (3) show how to choose metrics to produce never-before-seen seemingly emergent abilities in multiple vision tasks across diverse deep network architectures. Via all three analyses, we provide evidence that emergent abilities disappear with different metrics or with better statistics, and may not be a fundamental property of scaling AI models.

Figure 1: **Emergent abilities of large language models**. Model families display *sharp* and *unpredictable* increases in performance at specific tasks as scale increases. Source: Fig. 2 from [38].

Emergence in LLMs?

Hypothesis: What really causes emergence?

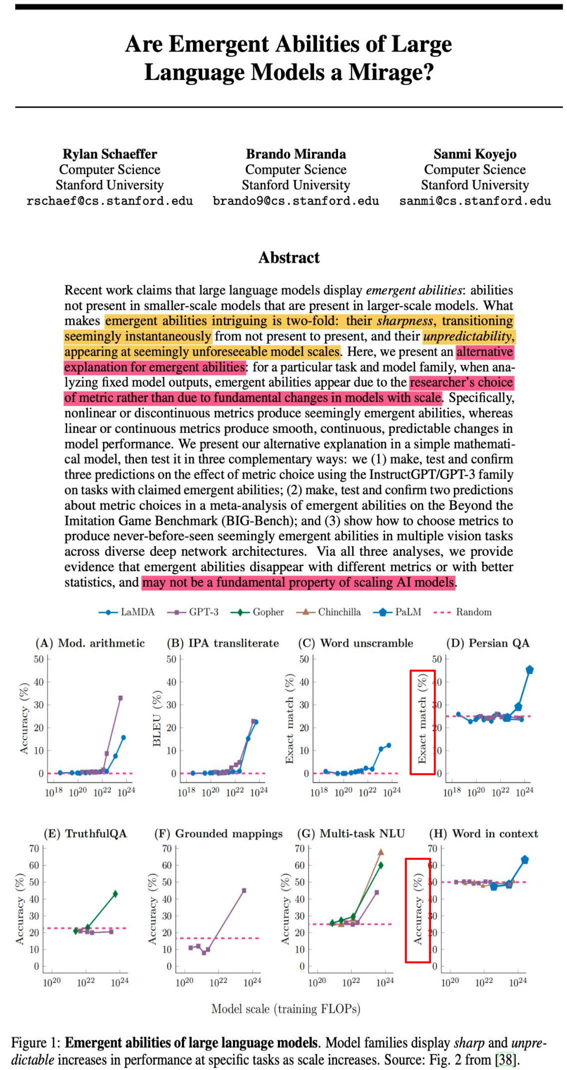
Which vector to choose?

1. Is it due to model scoring metric?
2. Is it due to low number of sample in test set?
3. Is it due to sparse sampling of model scales?

1. Simplest to change + most common pattern in y-axis

2. Some tasks are hard to create more data (Persian QA)

3. Very Expensive to do anywhere



Are Emergent Abilities of Large Language Models a Mirage?

Rylan Schaeffer
Computer Science
Stanford University
rschaef@cs.stanford.edu

Brando Miranda
Computer Science
Stanford University
brando9@cs.stanford.edu

Sanmi Koyejo
Computer Science
Stanford University
sanmi@cs.stanford.edu

Abstract

Recent work claims that large language models display *emergent abilities*: abilities not present in smaller-scale models that are present in larger-scale models. What makes emergent abilities intriguing is two-fold: their *sharpness*, transitioning seemingly instantaneously from not present to present, and their *unpredictability*, appearing at seemingly unforeseeable model scales. Here, we present an alternative explanation for emergent abilities: for a particular task and model family, when analyzing fixed model outputs, emergent abilities appear due to the researcher's choice of metric rather than due to fundamental changes in models with scale. Specifically, nonlinear or discontinuous metrics produce seemingly emergent abilities, whereas linear or continuous metrics produce smooth, continuous, predictable changes in model performance. We present our alternative explanation in a simple mathematical model, then test it in three complementary ways: we (1) make, test and confirm three predictions on the effect of metric choice using the InstructGPT/GPT-3 family on tasks with claimed emergent abilities; (2) make, test and confirm two predictions about metric choices in a meta-analysis of emergent abilities on the Beyond the Imitation Game Benchmark (BIG-Bench); and (3) show how to choose metrics to produce never-before-seen seemingly emergent abilities in multiple vision tasks across diverse deep network architectures. Via all three analyses, we provide evidence that emergent abilities disappear with different metrics or with better statistics, and may not be a fundamental property of scaling AI models.

Figure 1: **Emergent abilities of large language models.** Model families display *sharp* and *unpredictable* increases in performance at specific tasks as scale increases. Source: Fig. 2 from [38].

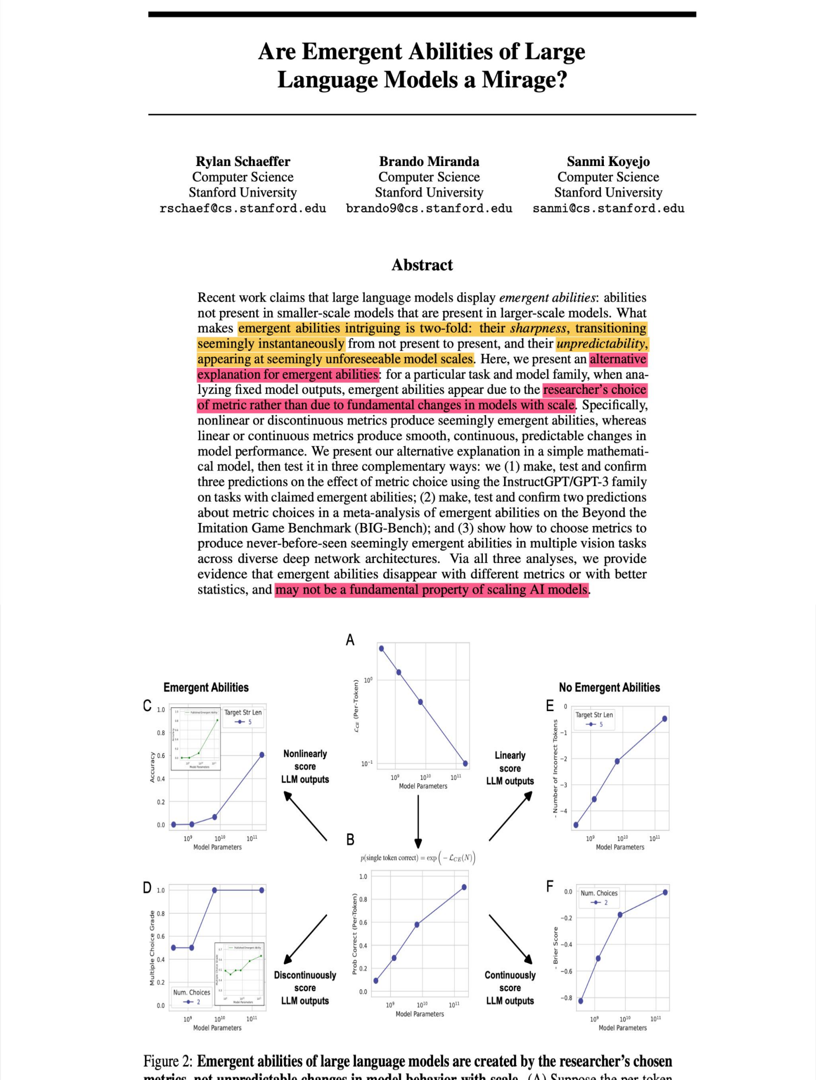
Emergence in LLMs?

Hypothesis: What really causes emergence?

Vector to chosen

1. Is it due to scoring metric?

Ultimately, that gave us experimental data very quickly in a direction of uncertainty/risk.



Are Emergent Abilities of Large Language Models a Mirage?

Rylan Schaeffer
Computer Science
Stanford University
rschaef@cs.stanford.edu

Brando Miranda
Computer Science
Stanford University
brando9@cs.stanford.edu

Sanmi Koyejo
Computer Science
Stanford University
sanmi@cs.stanford.edu

Abstract

Recent work claims that large language models display *emergent abilities*: abilities not present in smaller-scale models that are present in larger-scale models. What makes emergent abilities intriguing is two-fold: their *sharpness*, transitioning seemingly instantaneously from not present to present, and their *unpredictability*, appearing at seemingly unforeseeable model scales. Here, we present an alternative explanation for emergent abilities: for a particular task and model family, when analyzing fixed model outputs, emergent abilities appear due to the researcher's choice of metric rather than due to fundamental changes in models with scale. Specifically, nonlinear or discontinuous metrics produce seemingly emergent abilities, whereas linear or continuous metrics produce smooth, continuous, predictable changes in model performance. We present our alternative explanation in a simple mathematical model, then test it in three complementary ways: we (1) make, test and confirm three predictions on the effect of metric choice using the InstructGPT/GPT-3 family on tasks with claimed emergent abilities; (2) make, test and confirm two predictions about metric choices in a meta-analysis of emergent abilities on the Beyond the Imitation Game Benchmark (BIG-Bench); and (3) show how to choose metrics to produce never-before-seen seemingly emergent abilities in multiple vision tasks across diverse deep network architectures. Via all three analyses, we provide evidence that emergent abilities disappear with different metrics or with better statistics, and may not be a fundamental property of scaling AI models.

A
$\mathcal{L}_{CE}$ (Per-Token)
10^0
10^{-1}
Model Parameters
Emergent Abilities
C
Accuracy
Target Str Len
5
Published Emergent Ability
Nonlinearly score LLM outputs
No Emergent Abilities
E
- Number of Incorrect Tokens
Target Str Len
5
Linearly score LLM outputs
B
$p(\text{single token correct}) = \exp\left(-\mathcal{L}_{CE}(N)\right)$
Prob Correct (Per-Token)
D
Multiple Choice Grade
Num. Choices
2
Published Emergent Ability
Discontinuously score LLM outputs
Continuously score LLM outputs
F
- Brier Score
Num. Choices
2
10^9
10^{10}
10^{11}
Model Parameters

Figure 2: **Emergent abilities of large language models are created by the researcher's chosen metrics, not unpredictable changes in model behavior with scale.** (A) Suppose the per-token

Social Constitutions

While everyone thinks that human written constitutions are the only way to align AI models, we think that allowing a model to write its own constitutions based on observing people's interactions is better and more democratic.

What's our first step?

Social Contract AI: Aligning AI Assistants with Implicit Group Norms

Jan-Philipp Fränken, Sam Kwok†, Peixuan Ye†, Kanishk Gandhi

Dilip Arumugam, Jared Moore, Alex Tamkin

Tobias Gerstenberg, Noah D. Goodman
Stanford University
jphilipp@stanford.edu

Abstract

We explore the idea of aligning an AI assistant by inverting a model of users' (unknown) preferences from observed interactions. To validate our proposal, we run proof-of-concept simulations in the economic *ultimatum game*, formalizing user preferences as policies that guide the actions of simulated players. We find that the AI assistant accurately *aligns* its behavior to match standard policies from the economic literature (e.g., selfish, altruistic). However, the assistant's learned policies lack robustness and exhibit limited *generalization* in an out-of-distribution setting when confronted with a currency (e.g., grams of medicine) that was not included in the assistant's training distribution. Additionally, we find that when there is *inconsistency* in the relationship between language use and an unknown policy (e.g., an altruistic policy combined with rude language), the assistant's learning of the policy is slowed. Overall, our preliminary results suggest that developing simulation frameworks in which AI assistants need to infer preferences from diverse users can provide a valuable approach for studying practical alignment questions.[1]

[a] Explicit Group Norms: Constitutional AI
User
Constitution
Prioritize others' needs over your own when distributing resources.
User
AI
0:10
1:9
Accept
User
Accept
User
[b]
Inferring Group Norms: Social Contract AI
1:9
2:8
Accept
Accept
User
User
User
Social Contract
AI
Inferred Contract
Prioritize the wellbeing of others while maintaining a minimal personal gain.
AI
1:9
3:7
Accept
Accept
User
User

Figure 1: Illustration of Constitutional AI (CAI) and Social Contract AI (SCAI) in the ultimatum game [Harsanyi, 1961]. In the ultimatum game, one player (the proposer) proposes a division of a pot of money (e.g., $10) with another player (the responder). The proposer **offers** a share, and the responder can either **accept** or **reject** the offered share. If the responder accepts, the money is distributed as proposed; if they reject it, neither player receives anything. [a] CAI uses explicit group norms such as a *constitution* or content policy to guide the AI assistant. [b] SCAI inverts a model of users' preferences from observed interactions and uses the inferred *social contract* as guiding principle for the AI assistant.

1 Introduction

Developing scalable methods for effectively steering AI systems is a key challenge for alignment research [Bowman et al., 2022]. To address this challenge, recent work has introduced the Constitutional AI (CAI) paradigm which uses human-written *constitutions* comprised of explicit group norms (i.e., "do not be hateful") as guiding principles for AI assistants [see Fig. 1a; Bai et al., 2022b]. While these methods provide effective means to align AI assistants, they also face challenges. For example,

[1] Code and Prompts

Social Constitutions

Possible vectors:

Can a model follow a constitution reliably?

Can a model write principles based on observed interactions?

Can a model reliably revise its principles?

Social Contract AI: Aligning AI Assistants with Implicit Group Norms

Jan-Philipp Fränken, Sam Kwok[†], Peixuan Ye[†], Kanishk Gandhi

Dilip Arumugam, Jared Moore, Alex Tamkin

Tobias Gerstenberg, Noah D. Goodman
Stanford University
jphilipp@stanford.edu

Abstract

We explore the idea of aligning an AI assistant by inverting a model of users' (unknown) preferences from observed interactions. To validate our proposal, we run proof-of-concept simulations in the economic *ultimatum game*, formalizing user preferences as policies that guide the actions of simulated players. We find that the AI assistant accurately *aligns* its behavior to match standard policies from the economic literature (e.g., selfish, altruistic). However, the assistant's learned policies lack robustness and exhibit limited *generalization* in an out-of-distribution setting when confronted with a currency (e.g., grams of medicine) that was not included in the assistant's training distribution. Additionally, we find that when there is *inconsistency* in the relationship between language use and an unknown policy (e.g., an altruistic policy combined with rude language), the assistant's learning of the policy is slowed. Overall, our preliminary results suggest that developing simulation frameworks in which AI assistants need to infer preferences from diverse users can provide a valuable approach for studying practical alignment questions.[1]

[a] Explicit Group Norms: Constitutional AI
User
User
Constitution
Prioritize others' needs over your own when distributing resources.
AI
0:10
1:9
Accept
Accept
User
User
[b] Inferring Group Norms: Social Contract AI
1:9
2:8
Accept
Accept
User
User
User
Social Contract
AI
Inferred Contract
Prioritize the wellbeing of others while maintaining a minimal personal gain.
AI
1:9
3:7
Accept
Accept
User
User

Figure 1: Illustration of Constitutional AI (CAI) and Social Contract AI (SCAI) in the ultimatum game [Harsanyi, 1961]. In the ultimatum game, one player (the proposer) proposes a division of a pot of money (e.g., $10) with another player (the responder). The proposer **offers** a share, and the responder can either **accept** or **reject** the offered share. If the responder accepts, the money is distributed as proposed; if they reject it, neither player receives anything. [a] CAI uses explicit group norms such as a *constitution* or content policy to guide the AI assistant. [b] SCAI inverts a model of users' preferences from observed interactions and uses the inferred *social contract* as guiding principle for the AI assistant.

1 Introduction

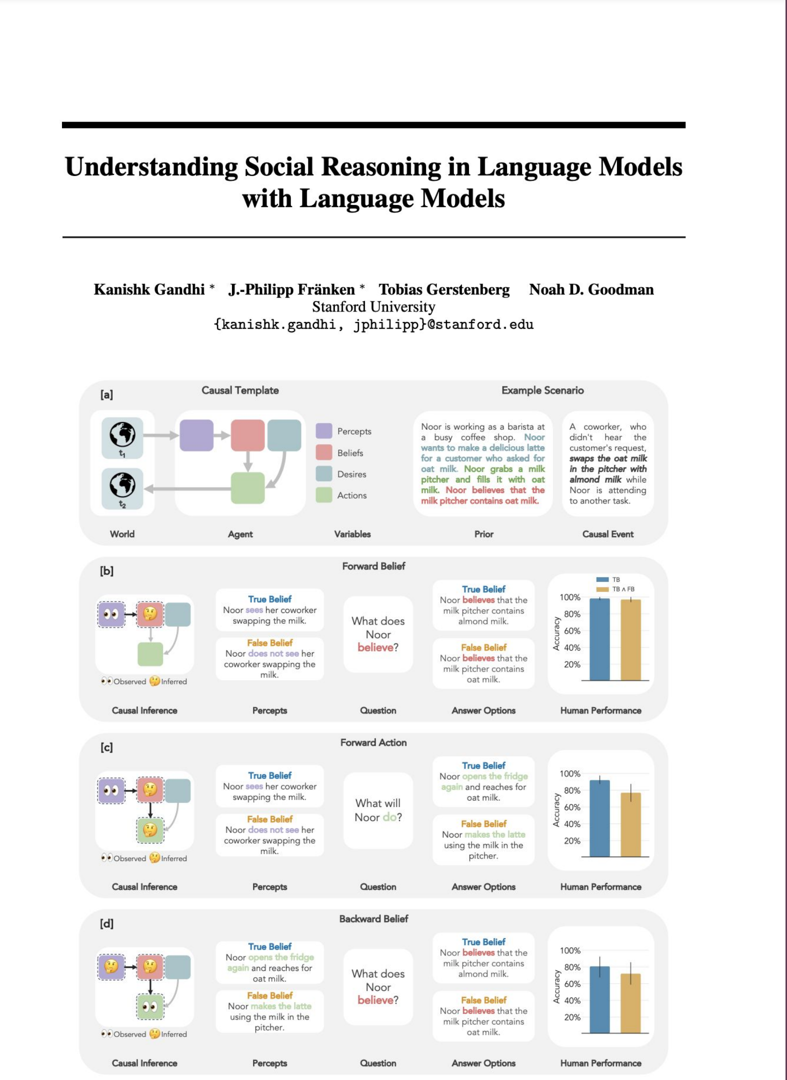
Developing scalable methods for effectively steering AI systems is a key challenge for alignment research [Bowman et al., 2022]. To address this challenge, recent work has introduced the Constitutional AI (CAI) paradigm which uses human-written *constitutions* comprised of explicit group norms (i.e., "do not be hateful") as guiding principles for AI assistants [see Fig. 1a; Bai et al., 2022b]. While these methods provide effective means to align AI assistants, they also face challenges. For example,

[1] Code and Prompts

Automated tests for reasoning

Everyone thought that we need to use tests for humans to test language models. We wanted to create a way to generate test for understanding social reasoning in language models with the use of language models themselves!

What’s our first step?



Understanding Social Reasoning in Language Models with Language Models

Kanishk Gandhi * J.-Philipp Fränken * Tobias Gerstenberg Noah D. Goodman
Stanford University
{kanishk.gandhi, jphilipp}@stanford.edu

Noor is working as a barista at a busy coffee shop. Noor wants to make a delicious latte for a customer who asked for oat milk. Noor grabs a milk pitcher and fills it with oat milk. Noor believes that the milk pitcher contains oat milk.

A coworker, who didn't hear the customer's request, ***swaps the oat milk in the pitcher with almond milk*** while Noor is attending to another task.

[b] Forward Belief

True Belief: Noor sees her coworker swapping the milk.

False Belief: Noor does not see her coworker swapping the milk.

What does Noor believe?

True Belief: Noor believes that the milk pitcher contains almond milk.

False Belief: Noor believes that the milk pitcher contains oat milk.

[c] Forward Action

True Belief: Noor sees her coworker swapping the milk.

False Belief: Noor does not see her coworker swapping the milk.

What will Noor do?

True Belief: Noor opens the fridge again and reaches for oat milk.

False Belief: Noor makes the latte using the milk in the pitcher.

[d] Backward Belief

True Belief: Noor opens the fridge again and reaches for oat milk.

False Belief: Noor makes the latte using the milk in the pitcher.

What does Noor believe?

True Belief: Noor believes that the milk pitcher contains almond milk.

False Belief: Noor believes that the milk pitcher contains oat milk.

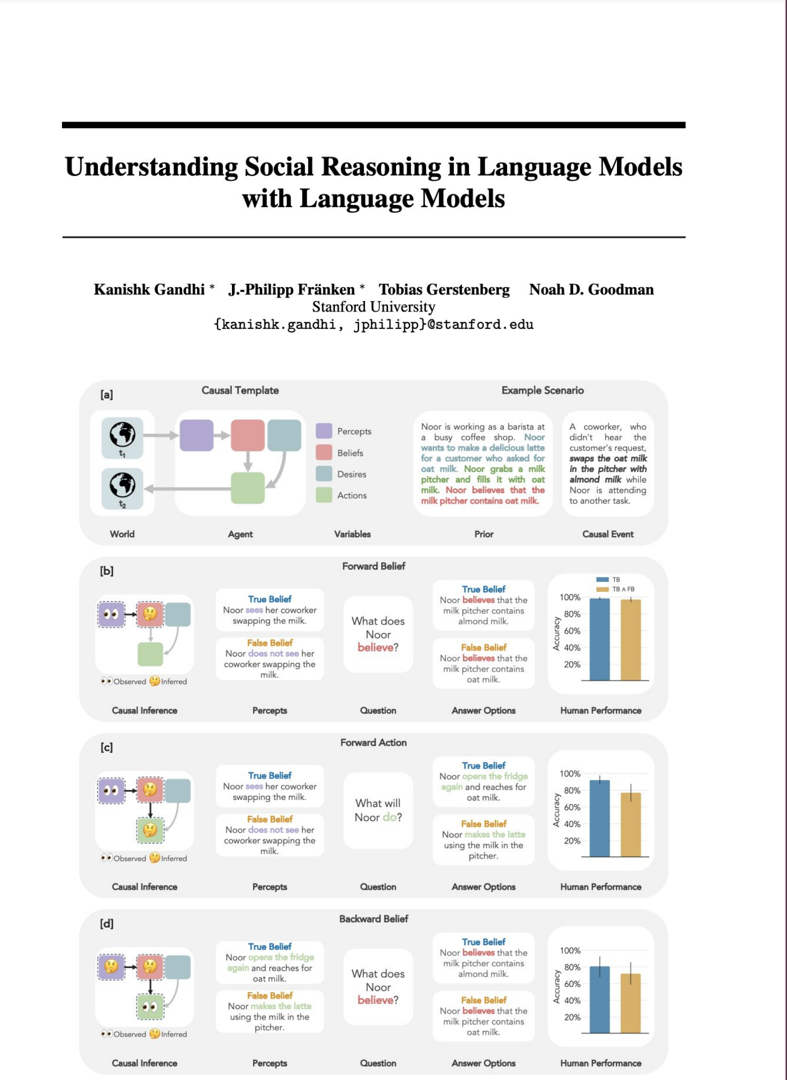
Automated tests for social reasoning

Possible vectors:

Can we create test items without having to reason through the test itself?

Can language models reliably follow instructions to generate this test?

If a model is generating its own test, can it be wrong while answering it?



Understanding Social Reasoning in Language Models with Language Models

Kanishk Gandhi * J.-Philipp Fränken * Tobias Gerstenberg Noah D. Goodman
Stanford University
{kanishk.gandhi, jphilipp}@stanford.edu

Noor is working as a barista at a busy coffee shop. Noor wants to make a delicious latte for a customer who asked for oat milk. Noor grabs a milk pitcher and fills it with oat milk. Noor believes that the milk pitcher contains oat milk.

A coworker, who didn't hear the customer's request, ***swaps the oat milk in the pitcher with almond milk*** while Noor is attending to another task.

True Belief
Noor sees her coworker swapping the milk.

False Belief
Noor does not see her coworker swapping the milk.

What does Noor believe?

True Belief
Noor believes that the milk pitcher contains almond milk.

False Belief
Noor believes that the milk pitcher contains oat milk.

True Belief
Noor sees her coworker swapping the milk.

False Belief
Noor does not see her coworker swapping the milk.

What will Noor do?

True Belief
Noor opens the fridge again and reaches for oat milk.

False Belief
Noor makes the latte using the milk in the pitcher.

True Belief
Noor opens the fridge again and reaches for oat milk.

False Belief
Noor makes the latte using the milk in the pitcher.

What does Noor believe?

True Belief
Noor believes that the milk pitcher contains almond milk.

False Belief
Noor believes that the milk pitcher contains oat milk.

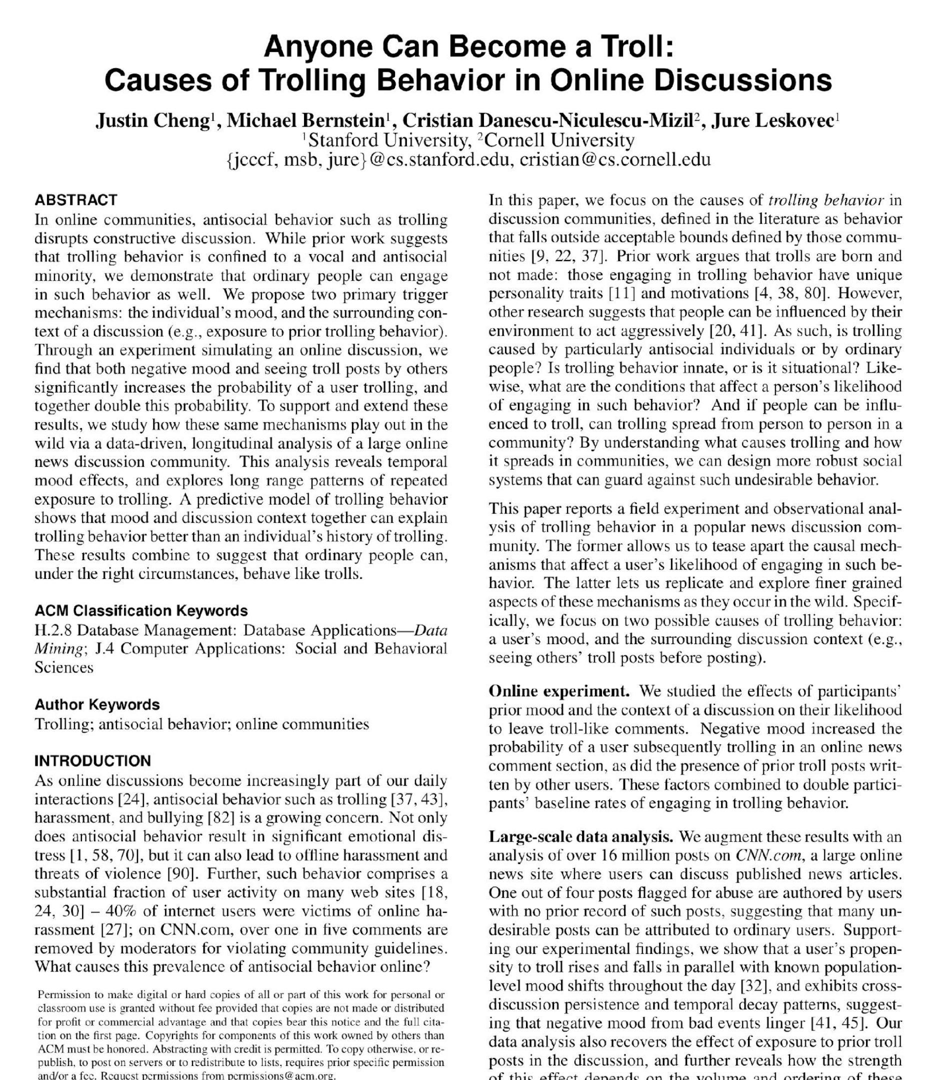
Trolling

Assumption: Everyone thinks that trolling online is due to a small number of antisocial sociopaths,

Hypothesis: we had a hunch that "normal" people were responsible for much trolling behavior when triggered.

What's our first step?

We have: dataset of 16M CNN comments (w/ troll flags), Mechanical Turk for studies



Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions

Justin Cheng[1], Michael Bernstein[1], Cristian Danescu-Niculescu-Mizil[2], Jure Leskovec[1]
[1]Stanford University, [2]Cornell University
{jcccf, msb, jure}@cs.stanford.edu, cristian@cs.cornell.edu

ABSTRACT

In online communities, antisocial behavior such as trolling disrupts constructive discussion. While prior work suggests that trolling behavior is confined to a vocal and antisocial minority, we demonstrate that ordinary people can engage in such behavior as well. We propose two primary trigger mechanisms: the individual's mood, and the surrounding context of a discussion (e.g., exposure to prior trolling behavior). Through an experiment simulating an online discussion, we find that both negative mood and seeing troll posts by others significantly increases the probability of a user trolling, and together double this probability. To support and extend these results, we study how these same mechanisms play out in the wild via a data-driven, longitudinal analysis of a large online news discussion community. This analysis reveals temporal mood effects, and explores long range patterns of repeated exposure to trolling. A predictive model of trolling behavior shows that mood and discussion context together can explain trolling behavior better than an individual's history of trolling. These results combine to suggest that ordinary people can, under the right circumstances, behave like trolls.

ACM Classification Keywords
H.2.8 Database Management: Database Applications—*Data Mining*; J.4 Computer Applications: Social and Behavioral Sciences

Author Keywords
Trolling; antisocial behavior; online communities

INTRODUCTION

As online discussions become increasingly part of our daily interactions [24], antisocial behavior such as trolling [37, 43], harassment, and bullying [82] is a growing concern. Not only does antisocial behavior result in significant emotional distress [1, 58, 70], but it can also lead to offline harassment and threats of violence [90]. Further, such behavior comprises a substantial fraction of user activity on many web sites [18, 24, 30] – 40% of internet users were victims of online harassment [27]; on CNN.com, over one in five comments are removed by moderators for violating community guidelines. What causes this prevalence of antisocial behavior online?

In this paper, we focus on the causes of *trolling behavior* in discussion communities, defined in the literature as behavior that falls outside acceptable bounds defined by those communities [9, 22, 37]. Prior work argues that trolls are born and not made: those engaging in trolling behavior have unique personality traits [11] and motivations [4, 38, 80]. However, other research suggests that people can be influenced by their environment to act aggressively [20, 41]. As such, is trolling caused by particularly antisocial individuals or by ordinary people? Is trolling behavior innate, or is it situational? Likewise, what are the conditions that affect a person's likelihood of engaging in such behavior? And if people can be influenced to troll, can trolling spread from person to person in a community? By understanding what causes trolling and how it spreads in communities, we can design more robust social systems that can guard against such undesirable behavior.

This paper reports a field experiment and observational analysis of trolling behavior in a popular news discussion community. The former allows us to tease apart the causal mechanisms that affect a user's likelihood of engaging in such behavior. The latter lets us replicate and explore finer grained aspects of these mechanisms as they occur in the wild. Specifically, we focus on two possible causes of trolling behavior: a user's mood, and the surrounding discussion context (e.g., seeing others' troll posts before posting).

Online experiment. We studied the effects of participants' prior mood and the context of a discussion on their likelihood to leave troll-like comments. Negative mood increased the probability of a user subsequently trolling in an online news comment section, as did the presence of prior troll posts written by other users. These factors combined to double participants' baseline rates of engaging in trolling behavior.

Large-scale data analysis. We augment these results with an analysis of over 16 million posts on *CNN.com*, a large online news site where users can discuss published news articles. One out of four posts flagged for abuse are authored by users with no prior record of such posts, suggesting that many undesirable posts can be attributed to ordinary users. Supporting our experimental findings, we show that a user's propensity to troll rises and falls in parallel with known population-level mood shifts throughout the day [32], and exhibits cross-discussion persistence and temporal decay patterns, suggesting that negative mood from bad events linger [41, 45]. Our data analysis also recovers the effect of exposure to prior troll posts in the discussion, and further reveals how the strength of this effect depends on the volume and ordering of these

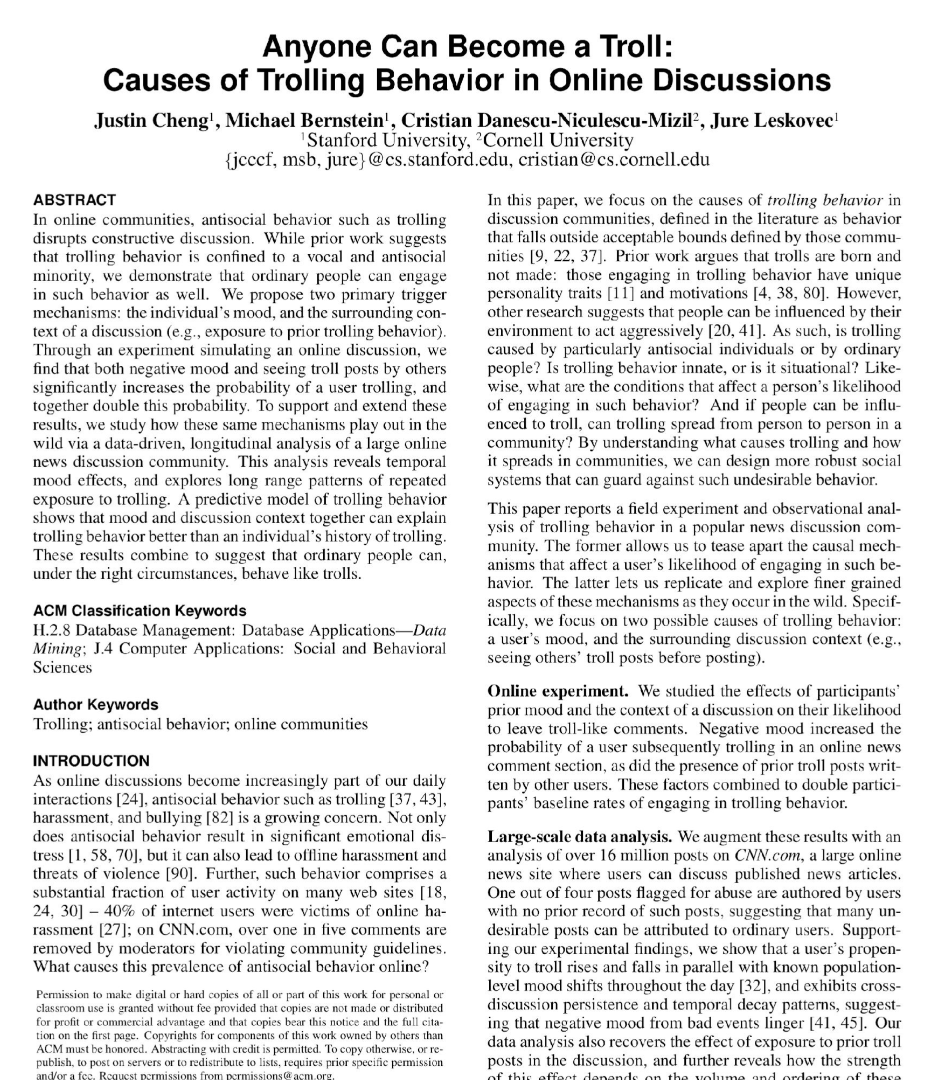
Trolling

Possible vectors:

1. Do people really troll when pissed off, check subset manually?

2. Can we train a classifier to predict when someone would troll, and compare weights of personal history vs. other posts and title?

3. Does the same person troll more on certain (angry) topics than on other (boring) ones?



Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions

Justin Cheng[1], **Michael Bernstein**[1], **Cristian Danescu-Niculescu-Mizil**[2], **Jure Leskovec**[1]
[1]Stanford University, [2]Cornell University
{jcccf, msb, jure}@cs.stanford.edu, cristian@cs.cornell.edu

ABSTRACT

In online communities, antisocial behavior such as trolling disrupts constructive discussion. While prior work suggests that trolling behavior is confined to a vocal and antisocial minority, we demonstrate that ordinary people can engage in such behavior as well. We propose two primary trigger mechanisms: the individual's mood, and the surrounding context of a discussion (e.g., exposure to prior trolling behavior). Through an experiment simulating an online discussion, we find that both negative mood and seeing troll posts by others significantly increases the probability of a user trolling, and together double this probability. To support and extend these results, we study how these same mechanisms play out in the wild via a data-driven, longitudinal analysis of a large online news discussion community. This analysis reveals temporal mood effects, and explores long range patterns of repeated exposure to trolling. A predictive model of trolling behavior shows that mood and discussion context together can explain trolling behavior better than an individual's history of trolling. These results combine to suggest that ordinary people can, under the right circumstances, behave like trolls.

ACM Classification Keywords

H.2.8 Database Management: Database Applications—*Data Mining*; J.4 Computer Applications: Social and Behavioral Sciences

Author Keywords

Trolling; antisocial behavior; online communities

INTRODUCTION

As online discussions become increasingly part of our daily interactions [24], antisocial behavior such as trolling [37, 43], harassment, and bullying [82] is a growing concern. Not only does antisocial behavior result in significant emotional distress [1, 58, 70], but it can also lead to offline harassment and threats of violence [90]. Further, such behavior comprises a substantial fraction of user activity on many web sites [18, 24, 30] – 40% of internet users were victims of online harassment [27]; on CNN.com, over one in five comments are removed by moderators for violating community guidelines. What causes this prevalence of antisocial behavior online?

In this paper, we focus on the causes of *trolling behavior* in discussion communities, defined in the literature as behavior that falls outside acceptable bounds defined by those communities [9, 22, 37]. Prior work argues that trolls are born and not made: those engaging in trolling behavior have unique personality traits [11] and motivations [4, 38, 80]. However, other research suggests that people can be influenced by their environment to act aggressively [20, 41]. As such, is trolling caused by particularly antisocial individuals or by ordinary people? Is trolling behavior innate, or is it situational? Likewise, what are the conditions that affect a person's likelihood of engaging in such behavior? And if people can be influenced to troll, can trolling spread from person to person in a community? By understanding what causes trolling and how it spreads in communities, we can design more robust social systems that can guard against such undesirable behavior.

This paper reports a field experiment and observational analysis of trolling behavior in a popular news discussion community. The former allows us to tease apart the causal mechanisms that affect a user's likelihood of engaging in such behavior. The latter lets us replicate and explore finer grained aspects of these mechanisms as they occur in the wild. Specifically, we focus on two possible causes of trolling behavior: a user's mood, and the surrounding discussion context (e.g., seeing others' troll posts before posting).

Online experiment. We studied the effects of participants' prior mood and the context of a discussion on their likelihood to leave troll-like comments. Negative mood increased the probability of a user subsequently trolling in an online news comment section, as did the presence of prior troll posts written by other users. These factors combined to double participants' baseline rates of engaging in trolling behavior.

Large-scale data analysis. We augment these results with an analysis of over 16 million posts on *CNN.com*, a large online news site where users can discuss published news articles. One out of four posts flagged for abuse are authored by users with no prior record of such posts, suggesting that many undesirable posts can be attributed to ordinary users. Supporting our experimental findings, we show that a user's propensity to troll rises and falls in parallel with known population-level mood shifts throughout the day [32], and exhibits cross-discussion persistence and temporal decay patterns, suggesting that negative mood from bad events linger [41, 45]. Our data analysis also recovers the effect of exposure to prior troll posts in the discussion, and further reveals how the strength of this effect depends on the volume and ordering of these

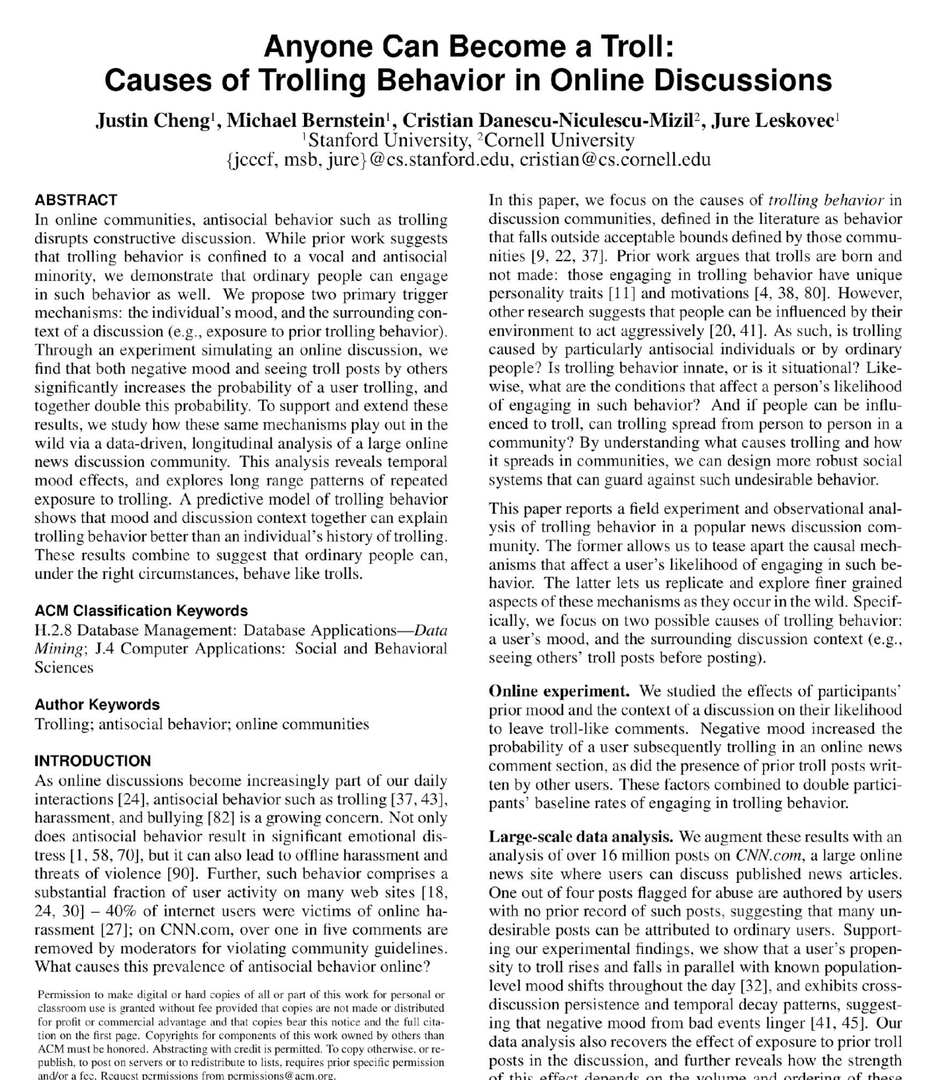
Trolling

Hypothesis: "normal" people troll when triggered.
Which vector to choose?

1. Do people really troll when pissed off, check subset manually?

2. Can we train a classifier to predict when someone would troll?

3. Does the same person troll more on certain (angry) topics than on other (boring) ones?



Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions

Justin Cheng[1], Michael Bernstein[1], Cristian Danescu-Niculescu-Mizil[2], Jure Leskovec[1]
[1]Stanford University, [2]Cornell University
{jcccf, msb, jure}@cs.stanford.edu, cristian@cs.cornell.edu

ABSTRACT

In online communities, antisocial behavior such as trolling disrupts constructive discussion. While prior work suggests that trolling behavior is confined to a vocal and antisocial minority, we demonstrate that ordinary people can engage in such behavior as well. We propose two primary trigger mechanisms: the individual's mood, and the surrounding context of a discussion (e.g., exposure to prior trolling behavior). Through an experiment simulating an online discussion, we find that both negative mood and seeing troll posts by others significantly increases the probability of a user trolling, and together double this probability. To support and extend these results, we study how these same mechanisms play out in the wild via a data-driven, longitudinal analysis of a large online news discussion community. This analysis reveals temporal mood effects, and explores long range patterns of repeated exposure to trolling. A predictive model of trolling behavior shows that mood and discussion context together can explain trolling behavior better than an individual's history of trolling. These results combine to suggest that ordinary people can, under the right circumstances, behave like trolls.

ACM Classification Keywords
H.2.8 Database Management: Database Applications—*Data Mining*; J.4 Computer Applications: Social and Behavioral Sciences

Author Keywords
Trolling; antisocial behavior; online communities

INTRODUCTION

As online discussions become increasingly part of our daily interactions [24], antisocial behavior such as trolling [37, 43], harassment, and bullying [82] is a growing concern. Not only does antisocial behavior result in significant emotional distress [1, 58, 70], but it can also lead to offline harassment and threats of violence [90]. Further, such behavior comprises a substantial fraction of user activity on many web sites [18, 24, 30] – 40% of internet users were victims of online harassment [27]; on CNN.com, over one in five comments are removed by moderators for violating community guidelines. What causes this prevalence of antisocial behavior online?

In this paper, we focus on the causes of *trolling behavior* in discussion communities, defined in the literature as behavior that falls outside acceptable bounds defined by those communities [9, 22, 37]. Prior work argues that trolls are born and not made: those engaging in trolling behavior have unique personality traits [11] and motivations [4, 38, 80]. However, other research suggests that people can be influenced by their environment to act aggressively [20, 41]. As such, is trolling caused by particularly antisocial individuals or by ordinary people? Is trolling behavior innate, or is it situational? Likewise, what are the conditions that affect a person's likelihood of engaging in such behavior? And if people can be influenced to troll, can trolling spread from person to person in a community? By understanding what causes trolling and how it spreads in communities, we can design more robust social systems that can guard against such undesirable behavior.

This paper reports a field experiment and observational analysis of trolling behavior in a popular news discussion community. The former allows us to tease apart the causal mechanisms that affect a user's likelihood of engaging in such behavior. The latter lets us replicate and explore finer grained aspects of these mechanisms as they occur in the wild. Specifically, we focus on two possible causes of trolling behavior: a user's mood, and the surrounding discussion context (e.g., seeing others' troll posts before posting).

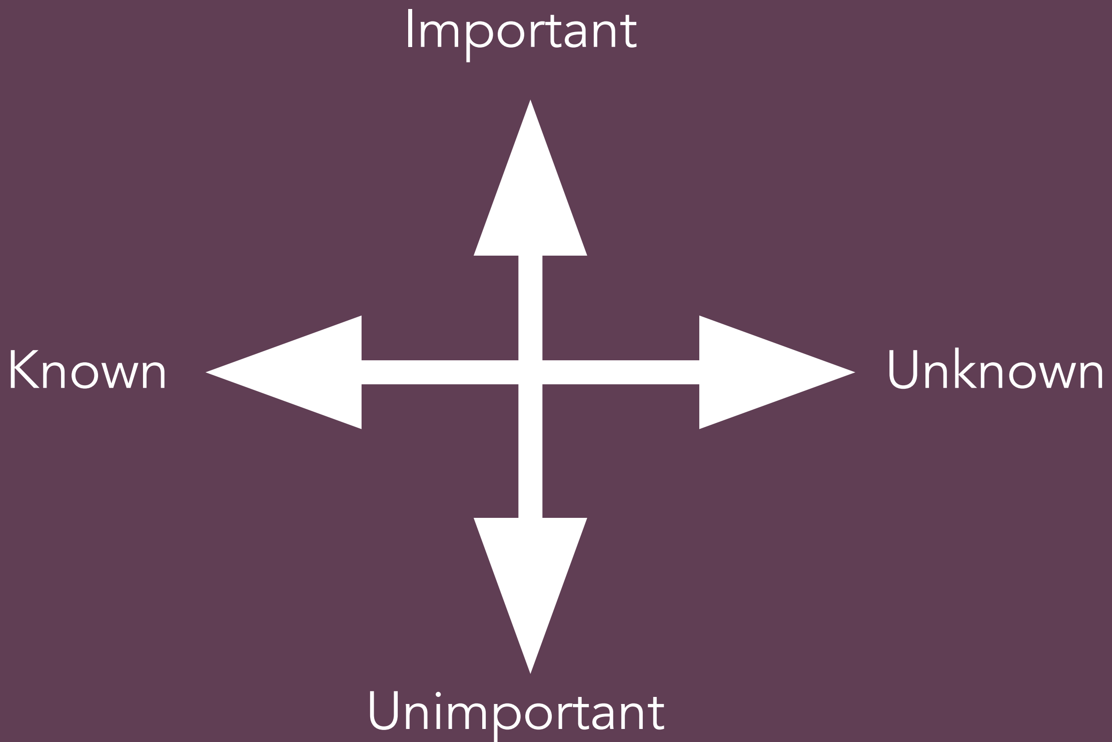
Online experiment. We studied the effects of participants' prior mood and the context of a discussion on their likelihood to leave troll-like comments. Negative mood increased the probability of a user subsequently trolling in an online news comment section, as did the presence of prior troll posts written by other users. These factors combined to double participants' baseline rates of engaging in trolling behavior.

Large-scale data analysis. We augment these results with an analysis of over 16 million posts on *CNN.com*, a large online news site where users can discuss published news articles. One out of four posts flagged for abuse are authored by users with no prior record of such posts, suggesting that many undesirable posts can be attributed to ordinary users. Supporting our experimental findings, we show that a user's propensity to troll rises and falls in parallel with known population-level mood shifts throughout the day [32], and exhibits cross-discussion persistence and temporal decay patterns, suggesting that negative mood from bad events linger [41, 45]. Our data analysis also recovers the effect of exposure to prior troll posts in the discussion, and further reveals how the strength of this effect depends on the volume and ordering of these

Assumption mapping

Assumption mapping is a strategy for articulating questions and ranking them.

Try assumption mapping your project [5min]



Tips and tricks

Vectoring and velocity

The output of a vectoring decision should allow you to identify what is core and what is periphery to reducing uncertainty in your vector of choice.

You should be able to make strong assumptions and use temporary scaffolding for anything that’s periphery.

(That’s the velocity skill.)

Why is vectoring so important?

"If Ernest Hemingway, James Mitchener, Neil Simon, Frank Lloyd Wright, and Pablo Picasso could not get it right the first time, what makes you think that you will?"

— Paul Heckel

Iteration >>> planning

Ideas rarely land exactly where you expect they will.

It’s best to test the most critical assumptions quickly, so that you can understand whether your hunch will play out, and what problems are worth spending time solving vs. kludging.

Human creative work is best in a loop of reflection and iteration.

Vectoring is a way to make sure you’re getting the most worthwhile iteration cycles – since vectoring attempt to choose the biggest uncertain direction idea working

Re-vectoring

Often, after vectoring and reducing uncertainty in one dimension, it raises new questions and uncertainties.

In the next round of vectoring, you re-prioritize:

If you get unexpected results and are confused (most of the time!), maybe it means you take a new angle to reduce uncertainty on a vector related to the prior one.

If you answer your question to your own satisfaction (not completely, just to your satisfaction), you move on to the next most important vector

Magnitude of your vector

The result of vectoring should be something achievable in about a week’s sprint. If it’s not, you’ve picked too broad a question to answer.

If you’re vectoring for “Can normal people be responsible for a lot of the trolling online?” is “Can normal people be responsible for a lot of the trolling on CNN.com?”, you’re still way too broad.

That’s evidence that you’ve just rescaled your project, not picked a vector.

Takeaways, in brief

1) The temptation is to try and solve the whole problem perfectly that’s set in front of you. *Don’t.*

2) Vectoring is a process of identifying the dimension of highest impact+uncertainty, and prioritizing that dimension while scaffolding the periphery

3) Successful vectoring enables you to rapidly hone in on the core insight of your research project

Progress Report++

At this point, your project transitions to a state where your team is working to try and achieve the goal you set out in Assignment 3.

Each week for the next several weeks, your team will perform vectoring, submit a brief summary, and report in section:

This week’s vector

This week’s plan

This week’s result

Vectoring in Research