ARO MURI: Foundations of Decision Making with Behavioral and Computational Constraints

https://aro-muri.lids.mit.edu/

Interim Report, August 2021

Ali Jadbabaie (lead PI), Elchanan Mossel, Josh Tenenbaum (MIT)
Austin Benson, Joe Halpern, Jon Kleinberg (Cornell)

1. Background and Introduction

Below we will outline our interim progress on the above named ARO MURI project over the interval of June 1, 2020 to August 15, 2021.

The goal of this multidisciplinary, multi-investigator research program is to create a rigorous foundation for individual and group decision-making under computational and cognitive constraints. The standard model of rational decision-making which has been the workhorse of decision theory and economic theory for 60-70 years maintains that rational individuals with limitless computational and cognitive abilities perform complex inferences use Bayes' rule to incorporate new information into their beliefs. In addition to its normative appeal, this rational paradigm serves as a highly useful benchmark by providing a well-grounded model.

Despite the success of this theory, a growing body of evidence over the past 2 has scrutinized this framework on the basis of its unrealistic cognitive demand on individuals and groups, especially when they make inferences in complex environments consisting of a large number of other decision-makers.

To address these issues, researchers have adopted an alternative paradigm by assuming non-Bayesian behavior of agents. Over the past two decades, the fields of behavioral economics and behavioral decision theory have sought to explain observed deviations from the predictions of rational decision making of individuals and groups. Despite some success in such efforts, the modeling approaches that result are typically ad hoc and fail to articulate what deviations from Bayesian rationality actually lead to the observed non-Bayesian behavior in the agents.

Consequently, we still lack a unified theory on human decision-making. Creating a foundational theory of behavioral and rational decision-making that addresses the above-mentioned shortcomings is of paramount importance to DoD in general, and the Army in particular.

Various vision documents for the Army of the future (2040 and beyond) have emphasized the importance of strategic command of information and decision-making in networks where individual agents have information of varying quality and precision, information exchange is limited and localized; decision making is purely local; and the sources, reliability, and trustworthiness of information is unclear. Existing theories of decision-making as the hand-crafted behavioral counterparts which are often qualitative in nature are wholly inadequate for dealing with these scenarios.

Our proposed effort is a step to fill these gaps and develop a better understanding of strengths and limitations of each approach in the individual and group settings. More importantly, we hope to bring ideas and principles from cognitive sciences to understand decision making processes subject to various types of cognitive limitations and biases in the same rigorous footing of rational decision-making theories.

To this end, our proposed effort is divided into three synergistic Thrusts, each composed of multiple interrelated tasks. In the first Thrust, we are working on development of a new set of theories for individual decision making under computational and cognitive constraints and develop hierarchy of models that rationalize different cognitive biases. In the second Thrust, we are working to develop a framework for the study of rational and behavioral decision making in groups, and more importantly to study the limitations of rational decision theory. In particular, we investigate how behavioral traits such as persuasion, polarization manifest through simple interactions among agents with various levels of rationality.

Our third Thrust is devoted to data-driven modelling of biases as well as behavioral and computational experiments. We have studied how certain habits and traits are formed and how a variety of behavioral biases can be studied using existing online datasets and further behavioral experiments.

Our team of 6 PIs consists of distinguished experts in a wide range of areas, from network science to computational social science and algorithms, decision theory, game theory, cognitive science, and collective behavior. All PIs have a track record in interdisciplinary research on the core topics of the proposal with a strong history of collaboration on previous and existing DoD efforts. Our distinguished team of PIs consists of laureates of MacArthur, Godel Prize, and Vannevar Bush ¨ Fellowships.

2. Project Thrusts and Tasks

For ease of collaboration and to ensure maximum positive externalities from collaboration between PIs and their researchers, graduate students and postdoctoral scholars, the project is decomposed into a series of thrusts and tasks as follows:

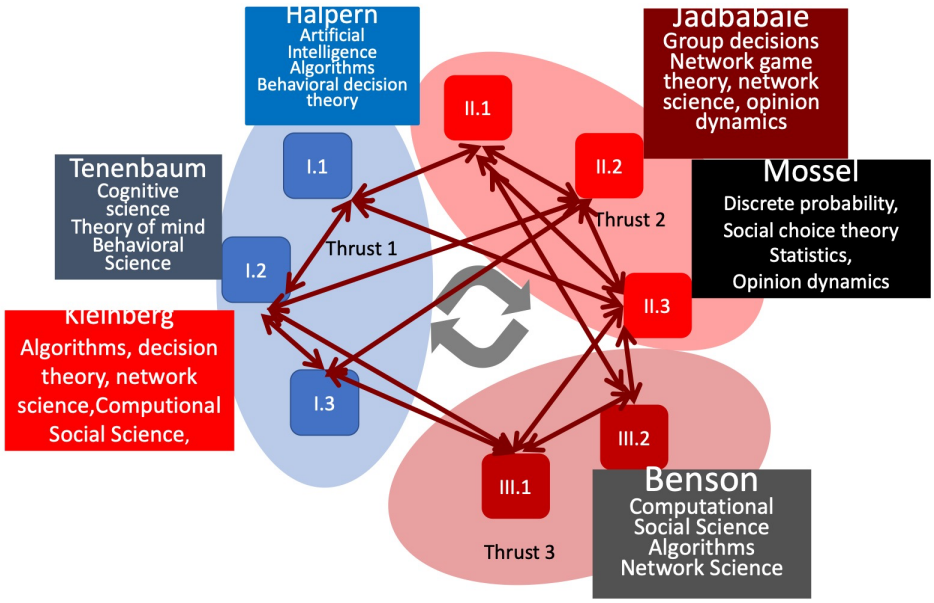
Thrust I: Individual Decision Making Under Constraints
Task [I.1]: Computationally-constrained Bayesian decisions
Task [I.2]: Choice-set models under computational constraints
Task [I.3]: Planning and decision making with biases

Thrust II: A Unified Theory of Group Decision Making
Task [II.1]: Group decisions and opinion exchange: role of persuasion
4.2 Task [II.2]: Group decision making: computational issues
4.3 Task [II.3]: Polarization
Thrust III: Modeling and Experimental Investigations
Task [III.1]: Data-driven study of cognitive biases

Task [III.2]: Online experiments on Amazon Mechanical Turk

3. Team, Research Thrusts, and Connections

The simple graph below shows the connections and synergies between tasks and thrusts of the project and the PI's expertise.



Graduate Students and Postdocs

The following students, postydocs and research staff contributed and were fully or partially supported on this project.

Cornell: Graduate students Meir Friedenberg, Shwan Ong, Oliver Richardson, and Spencer Peters, Kiran Tomilson, Ana Smith. Undergraduate student: Qian Huang

MIT: Research Scientist Amir Ajorlou. Postdoc: Yash Deshpande, Rabih Salhab, Sydney Levine, Farzan Farnia, Mh Tessler. Graduate students: Yan Jin, William Wang, Amir Tohidi, Jason Modeano.

4. Progress Report

Our efforts over the past year have lead to significant progress on all Thrusts of the project. Since the kick-off event in May 2019, our PIs have met regularly in large and small groups (weekly MURI meetings) first, to identify the type of talent we need to recruit for this project and then to attract them to the two institutions and get them involved in the project,

Below we will present some highlights and emphasize collaborations. Our joint efforts have led to joint publications in a diverse set of highly visible publication venues. More importantly we have seeded further substantive collaborations as well as follow-ups on other PIs works to make sure that the MURI project is indeed bigger than the sum of the parts, due to the positive externalities that the joint collaborations have fostered.
For simplicity, for each contribution, we list the publications corresponding to that work at the end of the same section.

[I1, I2, I3] Individual Decision Making Under Constraints with behavioral biases

Integrating explanation and prediction in computational social science

Computational social science is more than just large repositories of digital data and the computational methods needed to construct and analyse them. It also represents a convergence of different fields with different ways of thinking about and doing science. We aim to provide some clarity around how these approaches differ from one another and to propose how they might be productively integrated. Towards this end we make two contributions. The first is a schema for thinking about research activities along two dimensions—the extent to which work is explanatory, focusing on identifying and estimating causal effects, and the degree of consideration given to testing predictions of outcomes—and how these two priorities can complement, rather than compete with, one another. Our second contribution is to advocate that computational social scientists devote more attention to combining prediction and explanation, which we call integrative modelling, and to outline some practical suggestions for realizing this goal.

J. Hofman, D. Watts, S. Athey, F. Garip, T. Griffiths, J. Kleinberg, H. Margetts, S. Mullainathan, M. Salganik, S. Vazire, A. Vespignani, T. Yarkoni, "Integrating explanation and prediction in computational social science," *Nature* 595(2021).

Decision-Making and Moral Rules
Our work this year has focused on understanding *moral rules* as amortizations of more complex mental processes. We have focused on two phenomena that demonstrate this idea: 1) our ability to figure out how to make new rules for novel circumstances and 2) our ability to decide when the rules apply and when they can be broken.

1) Making New Rules: Though rules are a key feature of our moral decision-making, sometimes there just is no rule telling us how to behave. This is particularly true when novel situations arise – for instance, when a new morally-charged product (like an electric car) becomes available or a novel opportunity to help or harm others arises (like getting a new vaccine). Sometimes, we make a decision about how to deal with a novel circumstance, and then re-use that decision over and over in a rule-like manner. How do we do this? How do we determine what a good moral rule would be in novel contexts?

This question is particularly difficult in *collective action problems,* such as deciding whether to purchase a "virtuous" product, such as environmentally-friendly cleaning supplies or sustainably sourced cheese. After all, it seems as if one person's

purchase makes no difference to the global problems these products attempt to solve. However, in aggregate, massive good or bad effects can result when everyone chooses to act (or not act) a certain way. Our research shows that decisions in these sorts are guided by the logic of "universalization." Universalization asks us to consider the question, *what would happen if everyone felt free to do that?* In a recent paper, we formalize universalization as a computational model and test the model on a series of experiments with adults and children, predicting subject judgments with fine-grained accuracy (Levine, Kleiman-Weiner, Schulz, Tenenbaum & Cushman, 2020).

In addition to explaining judgments in collective action problems, universalization is a promising mental mechanism for generating new moral rules, because it demands that we consider the impacts of general adoption of a particular policy.

2) Breaking Rules*:* Sometimes, even when a clear rule governs the case at hand, it seems that the right thing to do is to *break* the rule. As a case study for rule-breaking, in ongoing work, we are studying how children and adults decide when it is OK to cut in line. For instance, it may be OK to cut in line at a deli to prevent someone from going into diabetic shock, but not if you're a bit caffeine deprived. Likewise, it may be OK to cut if you were given the wrong order and need it replaced but not if you simply prefer not to wait. Some instances of rule-breaking, therefore, can be explained by utility calculus; we sometimes break the rules when doing so would bring about a large benefit to the rule-breaker. Other times, we break the rules because we understand the *function* of the rule (e.g. distributing resources in a fair and orderly manner) and that breaking the rule actually instantiates the function rather than undermining it.
Other times, we decide it would be OK to break a rule using a "virtual bargaining" approach – considering what two people would agree to in a negotiation (Levine, Kleiman-Weiner, Chater, Cushman & Tenenbaum, working paper). If an agreement could be reached (that would be mutually beneficial to both parties), then it is permissible to violate the rule that was previously established. Virtual bargaining allows us to flexibly navigate novel situations when purely rule-based thinking would lead to suboptimal outcomes.

Many AI/ML-based algorithms designed to maximize preferences (e.g., for movie-recommendations) use rule-based constraints to adhere to ethical standards. In ongoing work with IBM Research (Awad, Kleiman-Weiner, Levine, et al., 2020), we aim to show that rule-based constraints are insufficient to mimic human-like moral judgment. We aim to design better preference-optimization algorithms based on our understanding of when people would find rule-breaking to be morally permissible.

Levine, S., Kleiman-Weiner, M., Schulz, L., Cushman, F. and Tenenbaum, J. (2020). The logic of universalization guides moral judgment. *Proceedings of the National Academy of Sciences*, 117(42), 26158-26169.https://www.pnas.org/content/117/42/26158

Awad, E., Kleiman-Weiner, M., Levine, S., Loreggia, A., Mattei, N., Rahwan, I., Rossi, F., Talamadupula, K. and Tenenbaum, J. (2020.) "When Is It Morally Acceptable to Break the Rules? A Preference-Based Approach." 12th *Multidisciplinary Workshop on*

Advances in Preference Handling (MPREF) held at the European Conference on Artificial Intelligence (ECAI).

Levine, S., Kleiman-Weiner, M., Chater, N., Cushman, F. and Tenenbaum, J. When is it OK to break the rules? (In preparation for *Management Science.)*

A Language-based Decision Theory

Standard decision theory assumes that a problem is described in terms of *states* and *outcomes*. The decision maker is choosing among *acts*: functions from states to outcomes. In complicated decision problems, constructing the appropriate state space and outcome space is difficult: What is the "right" state space and outcome space if we are deciding whether to pull out of Afghanistan? What if we raise the minimum wage to $15? Over the past year, Halpern and his group have defined a more flexible, realistic approach where acts are identified with formulas: ("pull out of Afghanistan"; "raise the minimum wage to $15"). In this setting, states and outcomes are not part of the description. Furthermore, Acts are underspecified. what else will happen if we raise the minimum wage to $15? Will workers lose their jobs? Will there be more automation? In this setting, decision makers can reason about this using standard techniques of counterfactuals: What (do I believe) is the closest world to the current world where I raise the minimum wage? Are jobs lost in that world?

This work was summarized in the following paper:
A. Bjorndahl and J. Y. Halpern, "Language-based decisions, " *Proceedings of Eighteenth Conference on Theoretical Aspects of Rationality and Knowledge (TARK)*, 2021. https://arxiv.org/abs/2106.11494

Language, Communication and Decision-making

Our work this year has focused on understanding the (1) under-specification and context-sensitivity of human communication, (2) human and machine reasoning in logical scenarios, and (3) knowledge transmission processes that underly cultural learning in a novel video game setting.

In pursuit of (1), we investigate the computational underpinnings of "generic language". Generic language (e.g., "Birds have hollow bones") conveys generalizations about categories and is a fundamental and ubiquitous mechanism of learning about the world. How exactly this learning occurs, however, is unclear: Generics exhibit so much flexibility in how they are interpretated that a quantitative theory of how generics update beliefs has not only not been developed or tested but is often explicitly dismissed. In this work, we explore a hypothesis introduced by Tessler and Goodman (2019) that generics update beliefs via an uncertain threshold like a vague quantifier: A category–property generic (Ks F) means that the property F is expected to be relatively widespread in the category K, where what counts as being relatively widespread is a priori uncertain and resolved by consulting one's prior beliefs about the property. We compared this model to a family of alternatives with the same background knowledge but where generics have a

fixed, precise meaning; in addition, we formalized a quantitative model of a strictly conceptual-based approach to generics, where generics are a direct connection to conceptual knowledge. Across three experiments in which we both measure and manipulate prior beliefs, we found the uncertain meaning approach to generics to be the best explanation of participants' highly heterogeneous interpretations of novel generic statements. This result taken together with the result that the same model of generic meaning can explain the variability in generic endorsements as shown by Tessler & Goodman (2019) suggests that generics are an effective medium for faithful transmission of knowledge between interlocutors. This work adds to the growing enterprise of formal, quantitative studies of human understanding of generic language. This work is currently under revision at the journal Cognition (Tessler & Goodman, in under review).

In pursuit of (2), we examine human reasoning with syllogisms. Syllogistic reasoning lies at the intriguing intersection of natural and formal reasoning, of language and logic. Syllogisms comprise a formal system of reasoning yet make use of natural language quantifiers (e.g., "all", "some") and invite natural language conclusions. The conclusions people tend to draw from syllogisms deviate substantially from a purely logical perspective. In this work we ask: Are principles of natural language understanding to blame? We developed probabilistic pragmatics models of syllogistic reasoning, couched within the Rational Speech Act modelng framework, and explore the pressures that pragmatic reasoning place on the production of conclusions. We tested our models on a recent, large data set of syllogistic reasoning and find that reasoners select conclusions in a pragmatic manner with the goal of aligning the beliefs of a naive listener to those of their own. We compared our model to previously published models that implement two alternative theories – Mental Models and Probability Heuristics – finding that our model accounts for the full distributions of responses in a manner slightly better than previous accounts but with an order-of-magnitude fewer parameters. Our work introduces a view of human syllogistic reasoning as natural communication. This work is currently in revision for a special issues of the topiCS (Topics in Cognitive Science) journal (Tessler, Tenenbaum, & Goodman, in revision).

Human reasoning can often be understood as an interplay between two systems: the intuitive and associative ("System 1") and the deliberative and logical ("System 2"). Neural sequence models—which have been increasingly successful at performing complex, structured tasks—exhibit the advantages and failure modes of System 1: they are fast and learn patterns from data but are often inconsistent and incoherent. In this work, we seek a lightweight, training-free means of improving existing System 1-like sequence models by adding System 2-inspired logical reasoning. We explore several variations on this theme in which candidate generations from a neural sequence model are examined for logical consistency by a symbolic reasoning module, which can either accept or reject the generations. Our approach uses neural inference to mediate between the neural System 1 and the logical System 2. Results in robust story generation and grounded instruction-following show that this approach can increase the coherence and accuracy of neural-based generations. This work is currently under review at the NeurIPS conference proceedings (Nye, Tessler, Tenenbaum, & Lake, 2021).

In pursuit of (3), we investigate human cultural learning in novel video games. Knowledge built culturally across generations allows humans to learn far more than an individual could glean from their own experience in a lifetime. Cultural knowledge in turn rests on language: language is the richest record of what previous generations believed, valued, and practiced. The power and mechanisms of language as a means of cultural learning, however, are not well understood. Our first project (Tessler, Tsividis, Madeano, Harper, & Tenenbaum, 2021) takes a step towards reverse-engineering cultural learning through language. We developed a suite of complex high-stakes tasks in the form of minimalist-style video games, which we deployed in an iterated learning paradigm. Game participants were limited to only two attempts (two lives) to beat each game and were allowed to write a message to a future participant who read the message before playing. We found that knowledge accumulated gradually across generations, allowing later generations to advance further in the games and perform more efficient actions. Multigenerational learning followed a strikingly similar trajectory to individuals learning alone with an unlimited number of lives. These results suggest that language provides a sufficient medium to express and accumulate the knowledge people acquire in these diverse tasks: the dynamics of the environment, valuable goals, dangerous risks, and strategies for success. The video game paradigm we pioneer here is thus a rich test bed for theories of cultural transmission and learning from language.

Tessler, M. H., Tsividis, P. A., Madeano, J., Harper, B., & Tenenbaum, J. B. (2021). Growing knowledge culturally across generations to solve novel, complex tasks. arXiv preprint arXiv:2107.13377.

Tessler, M. H., & Goodman, N. D. (under review). Learning from generic language. Cognition.

Tessler, M. H., Tenenbaum, J.B., & Goodman, N. D. (under review). Logic, probability, and pragmatics in syllogistic reasoning. topiCS.

Nye, M., Tessler, M. H., Tenenbaum, J. B., & Lake, B. M. (2021). Improving Coherence and Consistency in Neural Sequence Models with Dual-System, Neuro-Symbolic Reasoning. arXiv preprint arXiv:2107.02794.

Probabilistic Dependency Graphs

Probabilistic graphical models, such as Bayesian networks, are a key tool for representing and reasoning about probabilistic information. Halpern and his Ph.D. student Oliver Richardson introduced *Probabilistic Dependency Graphs (PDGs)*, a new class of directed graphical models. PDGs can capture inconsistent beliefs in a natural way, e.g., we can represent inconsistent reports from different sources. PDGs are more modular than Bayesian Networks (BNs); it is easier to incorporate new information, and Bayesian networks can be viewed as a special case of PDGs. Furthermore, inference in a PDG can be viewed as a process of reducing inconsistency

O. Richardson and J. Y. Halpern, "Probabilistic dependency graphs, "*Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)*, 2021. https://arxiv.org/abs/2012.10800

Dynamic Partial Awareness

Agents must sometimes take actions in situations that they do not fully comprehend. Work in computer science, economics, and philosophy has sought to capture this formally by considering agents who are unaware of some aspects of the world. Most work on awareness thus far has focused on the static case, where awareness does not change.

The focus of the recent work of Halpern and his student is the dynamic case: how to how to model the beliefs of an agent who becomes more aware. When there is no unawareness, the standard approach to dealing with beliefs is well understood; we update by conditioning. However, it is far from clear how beliefs should change when an agent becomes aware of new features. For example, a mathematician who becomes aware of the Riemann hypothesis, but learns nothing beyond that, may change his methods of proof and his beliefs about what is true, despite not having learned that any particular event has obtained. When an agent becomes more aware, his entire (subjective) view may change. In this work, PI Halpern and his student propose a model for the dynamics of increased awareness for agents who are introspective, and so can reason about their own and other's awareness (and lack of it). They work with a probabilistic extension of the framework of Halpern and Rego (2013) where agents understand that they themselves may be unaware of some propositions but cannot directly reason about or articulate such propositions. Instead, the agents consider states (possible worlds) that contain shadow propositions—proxies that represent an agent's vague conception of those propositions of which he is unaware. The states of the updated model resemble the original states except that some shadow propositions may be replaced by the novel propositions contained in the newly discovered formula.

This work has been summarized in the following papers:
J. Y. Halpern and E. Piermont, "Dynamic partial awareness, "
Proceedings of the Seventeenth International Conference on Principles of Knowledge Representation and Reasoning 2020. https://arxiv.org/abs/2007.02823

Decision-making and Behavioral Biases

Modeling Human Perceptions of Allocation Policies with Uncertain Outcomes: Many policies allocate harms or benefits that are uncertain in nature: they produce distributions over the population in which individuals have different probabilities of incurring harm or benefit. Comparing different policies thus involves a comparison of their corresponding probability distributions, and we observe that in many instances the policies selected in practice are hard to explain by preferences based only on the expected value of the total harm or benefit they produce. In cases where the expected value

analysis is not a sufficient explanatory framework, what would be a reasonable model for societal preferences over these distributions? Here we investigate explanations based on the framework of probability weighting from the behavioral sciences, which over several decades has identified systematic biases in how people perceive probabilities. We show that probability weighting can be used to make predictions about preferences over probabilistic distributions of harm and benefit that function quite differently from expected-value analysis, and in a number of cases provide potential explanations for policy preferences that appear hard to motivate by other means. In particular, we identify optimal policies for minimizing perceived total harm and maximizing perceived total benefit that take the distorting effects of probability weighting into account, and we discuss a number of real-world policies that resemble such allocational strategies. Our analysis does not provide specific recommendations for policy choices, but is instead fundamentally interpretive in nature, seeking to describe observed phenomena in policy choices.

This work was performed in PI Kleinberg's group and the lead author, Hoda Heidari, is a former student of PI Jadbabaie who was a postdoc with PI Kleinberg and is now a faculty at CMU School of Computing. (Incidentally, she was supported on a previous ARO MURI with Jadbabaie as a lead PI).

H. Heidari, S. Barocas, J. Kleinberg, K. Levy "On Modeling Human Perceptions of Allocation Policies with Uncertain Outcomes,"*Proc. 22nd ACM Conference on Economics and Computation (EC), 2021*.

Optimal Stopping with Behaviorally Biased Agents:

The Role of Loss Aversion and Changing Reference Points: People are often reluctant to sell a house, or shares of stock, below the price at which they originally bought it. While this is generally not consistent with rational utility maximization, it does reflect two strong empirical regularities that are central to the behavioral science of human decision-making: a tendency to evaluate outcomes relative to a reference point determined by context (in this case the original purchase price), and the phenomenon of loss aversion in which people are particularly prone to avoid outcomes below the reference point. Here we explore the implications of reference points and loss aversion in optimal stopping problems, where people evaluate a sequence of options in one pass, either accepting the option and stopping the search or giving up on the option forever. The best option seen so far sets a reference point that shifts as the search progresses, and a biased decision-maker's utility incurs an additional penalty when they accept a later option that is below this reference point. We formulate and study a behaviorally well-motivated version of the optimal stopping problem that incorporates these notions of reference dependence and loss aversion. We obtain tight bounds on the performance of a biased agent in this model relative to the best option obtainable in retrospect (a type of prophet inequality for biased agents), as well as tight bounds on the ratio between the performance of a biased agent and the performance of a rational one. We further establish basic monotonicity results,

and show an exponential gap between the performance of a biased agent in a stopping problem with respect to a worst-case versus a random order. As part of this, we establish fundamental differences between optimal stopping problems for rational versus biased agents, and these differences inform our analysis.

J. Kleinberg, R. Kleinberg, S. Oren, "Optimal Stopping with Behaviorally Biased Agents: The Role of Loss Aversion and Changing Reference Points," *Proc. 22nd ACM Conference on Economics and Computation (EC), 2021*. https://arxiv.org/abs/2106.00604

Stochastic Model for Sunk Cost Bias

We present a novel model for capturing the behavior of an agent exhibiting sunk-cost bias in a stochastic environment. Agents exhibiting sunk-cost bias take into account the effort they have already spent on an endeavor when they evaluate whether to continue or abandon it. We model planning tasks in which an agent with this type of bias tries to reach a designated goal. Our model structures this problem as a type of Markov decision process: loosely speaking, the agent traverses a directed acyclic graph with probabilistic transitions, paying costs for its actions as it tries to reach a target node containing a specified reward. The agent's sunk cost bias is modeled by a cost that it incurs for abandoning the traversal: if the agent decides to stop traversing the graph, it incurs a cost proportional to the sum of costs it has already invested. We analyze the behavior of two types of agents: naive agents that are unaware of their bias, and sophisticated agents that are aware of it. Since optimal (bias-free) behavior in this problem can involve abandoning the traversal before reaching the goal, the bias exhibited by these types of agents can result in sub-optimal behavior by shifting their decisions about abandonment. We show that in contrast to optimal agents, it is computationally hard to compute the optimal policy for a sophisticated agent. Our main results quantify the loss exhibited by these two types of agents with respect to an optimal agent.

J. Kleinberg, S. Oren, M. Raghavan, N. Sklar, "Stochastic Model for Sunk Cost Bias," *Proc. 37th Conference on Uncertainty in Artificial Intelligence (UAI), 2021*. https://arxiv.org/abs/2106.11003

[II.1, II.2, II.3] Thrust II: A Unified Theory of Group Decision Making

Our team has made significant progress in this area by starting and continuing a set of joint projects that have evolved from out weekly MURI discussions between PIs and their research groups. Recall that the focus of this thrust is on group decision making, which has been the subject of intense work by multiple PIs over the past decade. Topics of interest here include understanding complexities of rational group decision making, a computational framework for analysis of group decision-making and opinion exchange problems under various constraints (such as summarization and coarsening of beliefs with language), a multiagent social choice theory, and more importantly a focus on effects of persuasion and affirmation seeking.

Below we will summarize the highlights:

Polarization in Geometric Opinion Dynamics: In light of increasing recent attention to political polarization, understanding how polarization can arise poses an important theoretical question. While more classical models of opinion dynamics seem poorly equipped to study this phenomenon, a recent novel approach by Hązła, Jin, Mossel, and Ramnarayan (HJMR) performed under this MURI in 2019, proposes a simple geometric model of opinion evolution (on which we reported last year, https://arxiv.org/abs/1910.05274) that provably exhibits strong polarization in specialized cases. Moreover, polarization arises quite organically in their model: in each time step, each agent updates opinions according to their correlation/response with an issue drawn at random. We further the study of polarization in related geometric models. We show that the exact form of polarization in such models is quite nuanced: even when strong polarization does not hold, it is possible for weaker notions of polarization to nonetheless attain. We provide a concrete example where weak polarization holds, but strong polarization provably fails. However, we show that strong polarization provably holds in many variants of the HJMR model, which are also robust to a wider array of distributions of random issues -- this indicates that the form of polarization introduced by HJMR is more universal than suggested by their special cases. We also show that the weaker notions connect more readily to the theory of Markov chains on general state spaces.

J. Gaitonde, J. Kleinberg, E. Tardos, "Polarization in Geometric Opinion Dynamics," *Proc. 22nd ACM Conference on Economics and Computation (EC), 2021.* https://arxiv.org/abs/2106.12459

Robust Judgement Aggregation and Complexity of Group Decision making

A function f : {0, 1} n → {0, 1} is called an approximate AND homomorphism if choosing x, y ∈ {0, 1} n uniformly at random, we have that f (x ∧ y) = f (x) ∧ f (y) with probability at least 1 − ε, where x ∧ y = (x1 ∧ y1, . . . , xn ∧ yn). We prove that if f : {0, 1} n → {0, 1} is an approximate AND-homomorphism, then f is δ-close to either a constant function or an AND function, where δ(ε) → 0 as ε → 0. This improves on a result of Nehama, who proved a similar statement in which δ depends on n. Our theorem implies a strong result on judgement aggregation in computational social choice. In the language of social choice, our result shows that if f isε-close to satisfying judgement aggregation, then it is δ(ε)-close to an oligarchy (the name for the AND function in social choice theory). This improves on Nehama's result, in which δ decays polynomially with n. Our result follows from a more general one, in which we characterize approximate solutions to the eigenvalue equation Tf = λд, where T is the downwards noise operator Tf (x) = Ey[f (x ∧ y)], f is [0, 1]-valued, and д is {0, 1}-valued. We identify all exact solutions to this equation, and show that any approximate solution in which Tf and λд are close is close to an exact solution.
Filmus Y, Lifshitz N, Minzer D, Mossel E, "AND testing and robust judgement aggregation," *In Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing* 2020 Jun 8 (pp. 222-233).

On hardness of rational group decision making

As we outlined in the Introduction, it is already well-understood and accepted that the computational and cognitive burdens of fully rational group decision making is beyond the capability of humans, as group sizes get large. In. fact our own simple experiments on Mechanical Turk have verified that human groups often suffer from correlation neglect and imperfect recall of the past. One key aspect of this difficulty is that it is very hard for human decision makers to disentangle which part of the information received from decision of social peers is new, and which part is already in the information set of each agent. But the key remining question has been how hard is really group decision making as a decision problem? We have finally been able to answer this question and close a gap in this body of knowledge that existed in the literature since the seminal works of Papadimitriou and Tsitsiklis in 1980s.

Specifically, we have studied the computations that rational agents undertake when exchanging opinions and decisions over a network. The agents repeatedly make decisions on their private information and take myopic actions that maximize their expected utility according to a fully rational posterior belief. We show that such computations are NP-hard for two natural utility functions: one with binary actions, and another where agents reveal their posterior beliefs. In fact, we show that distinguishing between posteriors that are concentrated on different states of the world is NP-hard. Therefore, even approximating the Bayesian posterior beliefs is hard. We also describe a natural search algorithm to compute agents' actions, which we call elimination of impossible signals, and show that if the network is transitive, the algorithm can be modified to run in polynomial time.

This model has been studied extensively in economics from seminal works of Aumann for two agents and later by others in 1970s and 1980s, but exact complexity of such a decision problem has been unknown. While it is known that the agents eventually agree with high probability on any network. We have shown that it is PSPACE-hard for the agents to compute their actions in this model. Furthermore, we show that it is equally difficult even to approximate an agent's posterior: It is PSPACE-hard to distinguish between the posterior being almost entirely concentrated on one state of the
world or another.

We have further refined the notions of equilibria that rational decision makers face in the above model. We have introduced the notion of social learning equilibria, a static equilibrium concept that abstracts away from the details of the given extensive form, but nevertheless captures the corresponding asymptotic equilibrium behavior. We establish general conditions for agreement, herding, and information aggregation in equilibrium, highlighting a connection between agreement and information aggregation.

- J. Hązła, A. Jadbabaie, E. Mossel, and M. A. Rahimian. Bayesian Decision Making in Groups is Hard, *Operations Research.* March 2021;69(2):632-54.https://pubsonline.informs.org/doi/abs/10.1287/opre.2020.2000

Inferring Corruption in Networks

We consider the problem of distributed corruption detection in networks. In this model each node of a directed graph is either truthful or corrupt. Each node reports the type (truthful or corrupt) of each of its out neighbors. If it is truthful, it reports the truth, whereas if it is corrupt, it reports adversarially. This model, first considered by Preparata, Metze and Chien in 1967, motivated by the desire to identify the faulty components of a digital system by having the other components checking them, became known as the PMC model. The main known results for this model characterize networks in which all corrupt (that is, faulty) nodes can be identified, when there is a known upper bound on their number. We are interested in networks in which a large fraction of the nodes can be classified. It is known that in the PMC model, in order to identify all corrupt nodes when their number is t, all in-degrees have to be at least t. In contrast, we show that in d regular-graphs with strong expansion properties, a $1-O(1/d)$ fraction of the corrupt nodes, and a $1-O(1/d)$ fraction of the truthful nodes can be identified, whenever there is a majority of truthful nodes. We also observe that if the graph is very far from being a good expander, namely, if the deletion of a small set of nodes splits the graph into small components, then no corruption detection is possible even if most of the nodes are truthful. Finally we discuss the algorithmic aspects and the computational hardness of the problem.

Alon, Noga, Elchanan Mossel, and Robin Pemantle. "Distributed corruption detection in networks." *Theory of Computing* 16, no. 1 (2020).
https://theoryofcomputing.org/articles/v016a001/

Opinion Dynamics Under Social Pressure

In this work, jointly administered by PIs Jadbabaie and Mossel, we investigate whether it is possible to infer true opinions of a group that is under social pressure, from their declared opinions. We present a novel model of this phenomena and characterize conditions under which "natural" estimators of true beliefs are strongly consistent. Our model consists of a network of agents that have static (true) binary beliefs. However the opinion they declare only equals true beliefs with *truth probabilities.* We show that this model gives rise to a self-reinforcing Pólya urn dynamics: Agent's *truth probability* at time *t* is proportional to total number of neighboring opinions that agree with its true belief until time *t-1*. Our main result indicates that If population does *not* contain large majority of true beliefs, then beliefs can be estimated by averaging declared opinions over time. Otherwise, beliefs cannot be estimated by averaging. The key idea behind this novel analysis is to analyze convergence of *truth probability* process using stochastic approximations

A. Jadbabaie, A. Makur, E. Mossel & R. Salhab, "Opinion Dynamics under Social Pressure," submitted for publication to *IEEE Transactions on Automatic Control, https://arxiv.org/abs/2104.11172*

Behavioral Group Decisions and Fluid Democracy

Fluid democracy is a voting paradigm that allows voters to choose between directly voting and transitively delegating their votes to other voters. While fluid democracy has been viewed as a system that can combine the best aspects of direct and representative democracy, it can also result in situations where few voters amass a large amount of influence. To analyze the impact of this shortcoming, we consider what has been called an epistemic setting, where voters decide on a binary issue for which there is a ground truth. Previous work has shown that under certain assumptions on the delegation mechanism, the concentration of power is so severe that fluid democracy is less likely to identify the ground truth than direct voting. We examine different, arguably more realistic, classes of mechanisms, and prove they behave well by ensuring that (with high probability) there is a limit on concentration of power. Our proofs demonstrate that delegations can be treated as stochastic processes and that they can be compared to well-known processes from the literature -- such as preferential attachment and multi-types branching process -- that are sufficiently bounded for our purposes. Our results suggest that the concerns raised about fluid democracy can be overcome, thereby bolstering the case for this emerging paradigm.

This work was performed by 3 PIS of the MURI together with Ariel Procaccia at Harvard. Daniel Halpern, Joseph Y. Halpern, Ali Jadbabaie, Elchanan Mossel, Ariel D. Procaccia, Manon Revel, " In Defense of Fluid Democracy," *submitted for publication, 2021*. https://arxiv.org/abs/2107.11868

Persuasion, news sharing and information cascades

We study a model of online news dissemination on a Twitter-like social network. Given a news item and its credibility, agents with heterogeneous priors strategically decide whether to share the news with their followers. An agent shares the news, if the news can persuade her followers to take an action (such as voting) in line with the agent's perspectives. We describe the agent's decision making and the conditions that lead to sharing the news with followers, and characterize the size of news spread at the equilibrium of the news-sharing game. We further investigate the impact of the network connectivity, heterogeneity of prior perspectives, and news credibility on the set of the news that can trigger a sharing cascade. Finally, we identify the conditions under which the news with low credibility can spread wider than highly credible news.

In particular, we show that when the network is highly-connected or the news is not a ``tail event", a sharing cascade can occur even with news that is not credible.

Chin-Chia Hsu, Amir Ajorlou, Ali Jadbabaie, "Persuasion, news sharing and information cascades," to appear, *Proceedings of the IEEE Conference on Decision and Control*, CDC 2021, December 2021.

Information Disclosure and network formation in news subscription

We study the formation of a subscription network where a continuum of strategic, Bayesian subscribers decide to subscribe to one of two sources (leaders) for news that is informative about an underlying state of the world. The leaders, aiming to maximize the welfare of all subscribers, have a motive to persuade the subscribers to take the optimal binary action against the state according to their own perspectives. With this persuasion motive, each leader decides whether to disclose the news to her own subscribers when there is news. When the subscribers receive the news, they update their beliefs; more importantly, even when no news is disclosed, the subscribers update their beliefs, speculating that there may be news that was concealed due to the leader's strategic disclosure decision.

We prove that at any equilibrium, the set of news signals that are concealed by the leaders takes the form of an interval. We further show that when two leaders represent polarized and opposing perspectives, anti-homophily emerges among the subscribers whose perspectives are in the middle. For any subscriber with a perspective on the extremes, and for any leader, there exists an equilibrium at which the subscriber would follow the leader. Our results shed light on how individuals swould seek information when information is private or costly to obtain, while considering the strategic disclosure by the news providers who are partisan and have a hidden motive to persuade their followers. *Joint work by Muhamet Yildiz, Chin-Chia Haus, Ali Jadbabaie, and Amir Ajorlou*

Thrust III: Modeling and Experimental Investigations

Choice Models, context Effects, and Habits

Over the past year, PIs Jadbabaie and Benson have been regularly collaborating on developing new choice models and context effects. In particular, we have worked on a paper on learning what we call ``feature context effects" from data, some of which was presented in the review meeting last October.

Within behavioral economics, context effect models tend to be engineered to describe very specific effects and are often only applied (if at all) to carefully controlled special-purpose datasets. Psychological research on context effects in follow a similar pattern while also introducing complex behavioral processes (for instance, time-varying attention) that are typically not estimable from general choice datasets.

We developed methods for the automatic discovery of a wide class of context effects from large, pre-existing, and disparate choice datasets.

The key advantage of our approach over the previous work discussed above is that we can take a choice dataset collected in any domain (possibly one that has already been collected passively), efficiently train a model, and directly interpret the learned parameters as intuitive context effects.

We performed an extensive analysis of choice datasets using our models, showing that statistically significant feature context effects occur and recovering intuitive effects.

For example, we find evidence that people pick more expensive hotels when their choice sets have high star ratings, that people offered more oily sushi show more aversion to oiliness, and that when deciding whose Facebook wall to post on, people care more about mutual connections when choosing from popular friends.

We further studied causal inference for determining context effects from discrete choice data. There are many models for individual preferences, but nearly all learning methods overlook how choice set assignment affects the data.

Often, the choice set itself is influenced by an individual's preferences; for instance, a consumer choosing a product from an online retailer is often presented with options from a recommender system that depend on information about the consumer's preferences. Ignoring these assignment mechanisms can lead to biased estimates of preferences, which we call ``choice set confounding."

To combat this, we adapted methods from causal inference to the discrete choice setting, such as inverse probability weighting (IPW), which let us better answer counterfactuals. We showed that these methods reduce confounding on models trained on hotel booking data, making the choice system more consistent with utility-maximization and making inferred parameters more plausible. For example, the confounded data overweighs the importance of price, since many users are shown hotels matching their preferences and select the cheapest one. Factors such as star rating would play a more important role in counterfactuals.

- Kiran Tomlinson and Austin R. Benson, "Learning Interpretable Feature Context Effects in Discrete Choice," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data (KDD),* 2021. Code release: https://github.com/tomlinsonk/feature-context-effects

- Kiran Tomlinson, Johan Ugander, and Austin R. Benson, ``Choice Set Confounding in Discrete Choice," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data (KDD), 2021.* Code release: https://github.com/tomlinsonk/choice-set-confounding

- Derek Lim and Austin R. Benson, "Expertise and Dynamics within Crowdsourced Musical Knowledge Curation: A Case Study of the Genius Platform," *Proceedings of International Conference on Web and Social Media (ICWSM), 2021.* Code release: https://github.com/cptq/genius-expertise

- Qian Huang, Horace He, Abhay Singh, Yan Zhang, Ser-Nam Lim, and Austin R. Benson, "Better Set Representations For Relational Reasoning," *Advances in*

Neural Information Processing Systems (NeurIPS), 2020. Code release: https://github.com/CUAI/BetterSetRepresentations

Habit formation

We have begun to formulate new modelling and experimental efforts related to choice set effects. In the kick-off PI Benson described the inefficiencies that habits may cause if one views choices in the sequential decision making context. For example, he showed that subscribers of digital good and services follow a habitual pattern as opposed to balancing exploration and exploitation. We have started to expand on this work by using a data driven approach to consumers' shopping habits.

Our goal is to measure the effect of habit formation on consumers' behavior, in particular, the widespread experience of shopping in a store. In-store shopping is a repetitive behavior happening in the same context, so it has a great potential of becoming habitual. Although in the short run habits are formed to help people achieve their goals, in the long run, they could lead to non-optimal behavior. Therefore, it is important to understand the extent to which habit formation influences consumers' behavior.
We hypothesize that the measured state dependence in the literature is a combination of brand loyalty and shopping habits, and we try to disentangle these two phenomena. We use store closures as the exogenous shock that disrupts part of the households' shopping behavior. The main idea is that in a new environment, consumers are more engaged in thoughtful/deliberative decision-making processes, driving them to explore some other options that are normally ignored in a familiar store. The use of this exogenous shock is crucial to have a valid causal identification strategy and avoid various endogenous factors that could lead people to explore new stores. Having a better understanding of the psychological mechanisms behind our decision-making can benefit both individuals and firms. On the consumer side, it can help us design more effective interventions to improve people's healthcare, e.g., by nudging them to choose healthier options. Also, it can have managerial implications such as optimal pricing strategies or allocations of goods inside stores.

The outcome of this research is summarized in a paper entitled " Effect of Habit formation on consumer behavior," coauthored by MURI student Amir Tohidi, Dean Eckles, and PI Ali Jadbabaie.

5. Awards, Honors, and Keynote talks by PIs (June 2020-August 2021)

- Jon Kleinberg and coauthors received the Exemplary Applied Modeling Award at ACM EC 2021 conference (for the Heidari-Barocas-Kleinberg-Levy paper listed above, On Modeling Human Perceptions of Allocation Policies with Uncertain Outcomes).
- Austin Benson received the NSF Career Award
- Ali Jadbabaie received the 2021 HSCC/CPS Week Test of time Award
- Josh Tenenbaum became a AAAS member in 2020
- Student/postdoc placements:

    - Manish Raghavan (Kleinberg student) to join MIT in fall 2021
    - Anuran Makur (Jadbabaie postdoc) to join Purdue ECE as a faculty in fall 2021

- Plenary talks by Kleinberg:
    - International Conference on Game Theory, July 2021.
    - Research highlight talk, STOC/TheoryFest, June 2021.
    - Intersectional Symposium on Computing and Society, January 2021.
    - Future of Ads and Commerce Technology Conference, October 2020.
    - IEEE International Conference on Data Science and Advanced Analytics (DSAA), October 2020.
    - Heidelberg Laureate Forum, September 2020.

- Outreach and professional service:

    - Kleinberg served on National Academies Computer Science and Telecommunications Board study on Responsible Computing Research: Ethics and Governance of Computing Research and its Applications.
    - Opportunity and Inclusive Growth Institute Advisory Board, Federal Reserve Bank of Minneapolis.
    - Fairness, Accountability, and Transparency Conference Steering Committee.
    - Mechanism Design for Social Good (MD4SG) Executive Committee.
    - Jadbabaie co-organized the 3rd Learning for Dynamics and Control Conference
    - Jadbabaie appointed Department Head of Civil and Environmental Engineering at MIT

6. Spending

Our project is on track to spend most of the increment that was delivered in early June. We may have a short fall of around 100-150K due to the fact that two postdocs left early.

7. COVID 19 Impact

We have tried to minimize the impact of the COVID 19 pandemic on the project's progress. As of now, the project is not affected in a significant way. Because of their expertise many of the PIs are also helping MIT and Cornell come up with plans to open up in a responsible way. As of May 10, all meetings of the MURI have been on Zoom.