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

Interim Report, June 14, 2020

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, 2019 to June 14, 2020.

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

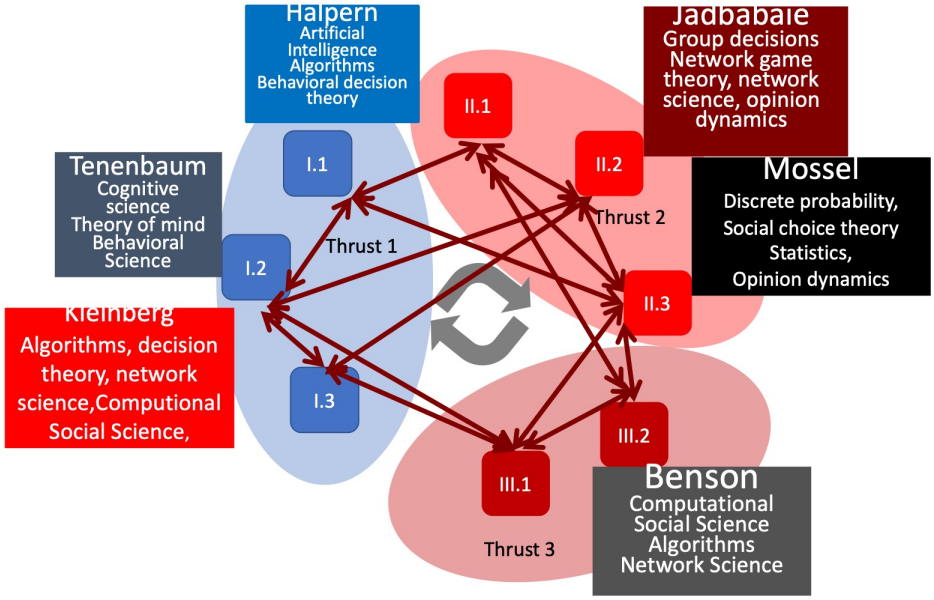
For ease of collaboration and to ensure maximum positive externalities from collaboration between PIs and their researchers, graduate students and postdoctoral scholars, the project is decomposed into a series of thrusts and tasks as follows:

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.



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

Explaining biases as decision making under constraints
As we stated in the Introduction, it is now well-understood that the idealized assumption of rational decision-making human agents who act to maximize a utility and have infinite foresight,

and computational and cognitive abilities is not supported by evidence. Below we describe some highlights of our advances in this area.

While traditional economics assumes that humans are fully rational agents who always maximize their expected utility, in practice, we constantly observe apparently irrational behavior. One explanation is that people have limited computational power, so that they are, quite rationally, making the best decisions they can, given their computational limitations. To test this hypothesis, Halpern and Lily Liu (a Cornell undergrad who will be pursuing a Ph.D. at MIT next year) consider the *multi-armed bandit (MAB) problem*. They examine a simple strategy for playing a MAB that maintains a relative rank of arms and essentially uses an elimination tournament to eliminate arms from consideration, while building in some reluctance to eliminate arms too quickly. The strategy can be implemented easily by a *probabilistic finite automaton* (PFA). When the PFA has sufficiently many states, it performs near-optimally. Its performance degrades gracefully as the number of states decreases. Moreover, the PFA exhibits *a negativity bias*, *a status quo bias*, and can be viewed as implementing an *availability heuristic*; that is, it exhibits human-like behavior.

- J. Y. Halpern and L. Liu, Bounded rationality in Las Vegas: probabilistic finite automata play multi-armed bandits, to appear, *Proceedings of the 36th Conference on Uncertainty in AI (UAI 2020)*, 2020.

Role of generics, statistical knowledge and observations in probabilistic language

Generic language (e.g., "Birds fly") conveys generalizations about categories and is essential for learning beyond our direct experience. The meaning of generic language is notoriously hard to specify, however (e.g., penguins don't fly). Recently Tessler and coauthor (Postdoc of CoPI Tenenbaum) proposed a model for generics that is mathematically equivalent to Bayesian belief-updating based on a single pedagogical example, suggesting a deep connection between learning from experience and learning from language. Relatedly, Csibra and Shamsudheen (2015) argue that generics are inherently pedagogical, understood by infants as referring to a member of a kind. In two experiments with adults, we quantify the exchange-rate between generics and observations by relating their belief-updating capacity, varying both the number of observations and whether they are presented pedagogically or incidentally. We find generics convey stronger generalizations than single pedagogical observations (Experiment 1), even when the property is explicitly demarcated (Experiment 2). We suggest revisions to the vague quantifier model of generics that would allow it to accommodate this intriguing exchange-rate.

Another hallmark of human reasoning is that we can bring to bear a diverse web of common-sense knowledge in any situation. The vastness of our knowledge poses a challenge for the practical implementation of reasoning systems as well as for our cognitive theories – how do people represent their common-sense knowledge? On the one hand, our best models of sophisticated reasoning are top-down, making use primarily of symbolically-encoded knowledge. On the other, much of our understanding of the statistical properties of our environment may arise in a bottom-up fashion, for example through associationist learning mechanisms. Indeed, recent advances in AI have

enabled the development of billion-parameter language models that can scour for patterns in gigabytes of text from the web, picking up a surprising amount of common-sense knowledge along the way—but they fail to learn the structure of coherent reasoning. We propose combining these approaches, by embedding language-model-backed primitives into a state- of-the-art probabilistic programming language (PPL). On two open-ended reasoning tasks, we show that our PPL models with neural knowledge components characterize the distribution of human responses more accurately than the neural language models alone, raising interesting questions about how people might use language as an interface to common-sense knowledge, and suggesting that building probabilistic models with neural language-model components may be a promising approach for more human-like AI.

- MH Tessler and Josh Tenenbaum, "How many observations is one generic worth?" *In proceedings of the 42nd Annual Conference of the Cognitive Science Society, CogSci 2020*

- MH Tessler and Josh Tenenbaum, "Leveraging Unstructured Statistical Knowledge in a Probabilistic Language of Thought," *In proceedings of the 42nd Annual Conference of the Cognitive Science Society, CogSci 2020*

Limited computation, computational games, and approximate equilibria

Over the past year, Halpern, Rafael Pass, and Daniel Reichman consider computational games, sequences of games G = ($G1$; $G2$; …) where, for all *n,* G_n has the same set of players. Computational games arise in electronic money systems such as Bitcoin, in cryptographic protocols, and in the study of generative adversarial networks in machine learning. Assuming that one-way functions exist, they prove that there is 2-player zero-sum computational game G such that, for all n, the size of the action space in G_n is polynomial in n and the utility function in G_n is computable in time polynomial in*n,* and yet there is no ε-Nash equilibrium if players are restricted to using strategies computable by polynomial-time Turing machines, where a notion of Nash equilibrium that is tailored to computational games is used. They also show that an ε-Nash equilibrium may not exist if players are constrained to perform at most T computational steps in each of the games in the sequence. On the other hand, they show that if players can use arbitrary Turing machines to compute their strategies, then every computational game has an ε-Nash equilibrium. These results may shed light on competitive settings where the availability of more running time or faster algorithms can lead to a "*computational arms race*", precluding the existence of equilibrium. They also point to inherent limitations of concepts such as \best response" and Nash equilibrium in games with resource-bounded players.

- J. Y. Halpern, R. Pass, and D. Reichman, On the existence of Nash Equilibrium in games with resource-bounded players, *Proceedings of the 12th International Symposium on Algorithmic Game Theory (SAGT)*, 2019, pp. 139--152.

Explaining cooperation among players with biases: non-utility maximizers
In the last few decades, numerous experiments have shown that humans do not always behave so as to maximize their material pay-offs. Cooperative behavior when non-cooperation is a dominant strategy (with respect to the material payoffs) is particularly puzzling: Over the past year, coPI Halpern and coauthor Valerio Capraro propose a novel approach to explain cooperation, assuming what Halpern and Pass have called translucent players. Typically, players are assumed to be opaque, in the sense that a deviation by one player in a normal-form game does not affect the strategies used by other players. But a player may believe that if he switches from one strategy to another, the fact that he chooses to switch may be visible to the other players. For example, if he chooses to defect in Prisoner's Dilemma, the other player may sense his guilt. They show that by assuming translucent players, we can recover many of the regularities observed in human behavior in well-studied games such as *Prisoner's Dilemma, Traveler's Dilemma, Bertrand Competition, and the Public Goods game*. The approach can also be extended to take into account a player's concerns that his social group (or God) may observe his actions. This extension helps explain prosocial behavior in situations in which previous models of social behavior fail to make correct predictions (e.g., conflict situations and situations where there is a tradeoff between equity and efficiency).

- V. Capraro and J. Y. Halpern, Translucent players: explaining cooperative behavior in social dilemmas*, Rationality and Society* **31**, 2019, pp. 371--408.

Emergence of moral decision making, contractualism, and universalization
What are the cognitive principles behind emergence of moral decision making? To explain why an action is wrong, we sometimes say: "What if everybody did that?" In other words, even if a single person's behavior is harmless, that behavior may be wrong if it would be harmful once universalized. We formalize the process of universalization in a computational model, test its quantitative predictions in studies of human moral judgment, and distinguish it from alternative models. We show that adults spontaneously make moral judgments consistent with the logic of universalization, and that children show a comparable pattern of judgment as early as 4 years old. We conclude that alongside other well-characterized mechanisms of moral judgment, such as outcome-based and rule-based thinking, the logic of universalizing holds an important place in our moral minds.

In addition, we have sketched out a computational account of an agreement-based model of moral decision-making. We first describe a series of experiments in which we ask participants to make judgments about when it is permissible to cut in line. We then describe a framework for modeling those judgments based on what the people in line would agree to. We point out how agreement is under-girded by a sophisticated utility calculus, which interacts with our understanding of the rules of waiting in line. Together, this model is the beginnings of a Psychological Triple Theory, an account by which agreement-based, outcome-based, and rule-based processes are all at play in moral decision-making.

Finally, we have studied the notion of *Contractualism* from a decision making and cognitive science perspective. Contractualism is a theory of moral philosophy that posits that an act is morally permissible if all the parties relevantly affected by the act could reasonably agree to it. We take this theory of moral philosophy as an inspiration for a theory of moral cognition. We have presented evidence that subjects have contractualist intuitions and use explicit contractualist reasoning. These data are poorly accounted for by current theories of moral cognition which rely mostly on the use of rules or calculations of consequences. We sketch out a rational model that captures these phenomena by predicting subjects' moral judgments as a function of their representation of the interests of agents who are engaged in a mentally simulated bargaining process. We have further provided a discussion on how a computational cognitive science of contractualism fits into a utility-based unified theory of moral cognition, which integrates elements of rule-based, consequence-based and contract-based cognitive mechanisms.

One puzzling aspect here is that there are some actions that if everyone did them there would be a terrible outcome. However, if just a few people did them, then utility would go up, at least for those people. Similar to the notion of what economists call public good games. An example is our over-fishing cases have this structure.

One thing to say about these actions is that everyone should be forbidden from doing them. (And our previous studies show that people judge that those actions are morally impermissible.) However, that is not the utility-maximizing solution we would come to if everyone involved could discuss the situation (i.e.: come to an actual agreement). The utility-maximizing solution is to use a more sophisticated policy that is actually universalizable. Are there ways for individuals to act on sustainable policies that are universalizable that reap the benefits of the new hook without risking disaster? What features would such a policy have?

Here are some further ideas on this topic that we are exploring for next year:

--The policy should be "legible" — others can figure out what you ae doing. You may personally be acting on a sustainable policy (e.g., only use the hook when I'm wearing my orange pants) but if no one can figure out that your policy is sustainable, that is problematic (both for your reputation but also because it makes it difficult for everyone else to condition their action on yours and make sure we're acting sustainably as a community).

--The policy should be "tamperproof" (as in the security literature): it is not easily game-able by someone who is trying to formally adhere to the policy but actually reap disproportionate rewards for himself (e.g. if the policy is "use the hook when your in-laws come to visit" I could start inviting my in-laws all the time).

--The policy should be "robust" (as in the game theory literature): if one person deviates (e.g. by accident) then the disastrous outcome still doesn't happen.

- Sydney Levine and Josh Tenenbaum, "Universalization reasoning guides moral judgment," *Proceedings of National Academy of Science (Revise and Resubmit)*

- Sydney Levine, Francesca Rossi, Josh Tenenbaum, "When is it OK to break the rules?," *12th Multidisciplinary Workshop on Advances in Preference Handling*

- Sydney Levine and Josh Tenenbaum, "The Cognitive Mechanisms of Contractualist Moral Decision-Making*," Journal paper in preparation*

Sequential Equilibria in Computational Games

Halpern and Pass also examined sequential equilibrium in the context of computational games, where agents are charged for computation. In such games, an agent can rationally choose to forget, so issues of imperfect recall arise. In this setting, they consider two notions of sequential equilibrium. One is an *ex ante* notion, where a player chooses his strategy before the game starts and is committed to it, but chooses it in such a way that it remains optimal even off the equilibrium path. The second is an *interim* notion, where a player can reconsider at each information set whether he is doing the "right" thing, and if not, can change his strategy. The two notions agree in games of perfect recall, but not in games of imperfect recall. Although the interim notion seems more appealing, in other work they have argued that there are some deep conceptual problems with it in standard games of imperfect recall. They show that the conceptual problems largely disappear in the computational setting. Moreover, in this setting, under natural assumptions, the two notions coincide.

- J. Y. Halpern and R. Pass, Sequential equilibrium in computational games, *ACM Transactions on Economics and Computation* 7:2, 2019.

Choice Set and context effects in human decision making

One high-level goal of the project is to better model and understand biases in human decision making. One particular method of influence is introducing new alternative choices for individual decisions. Early economic models assumed that alternatives are irrelevant to the relative ranking of options (e.g., the work of Luce in the 1950s and the Nobel-prize-winning research of McFadden in the 1970s), and experimental research has consistently found that new alternatives have strong effects on our choices. These effects are often called *context effects* or *choice set effects*. A well-known example is the compromise effect, which describes how people often prefer a middle ground.

In our recent research (*Tomlinson and Benson, ICML, 2020*), we considered adding new alter- natives as a discrete optimization problem for influencing a group of decision makers. In our setup, everyone has a base set of alternatives from which they make a choice and the goal is to introduce additional alternatives to optimize some function of the group's joint preferences on the base set. We specifically analyze three objective functions: (i) *agreement* in preferences amongst the group;

(ii) *disagreement* in preferences amongst the group; and (iii) *promotion* of a particular item.

We used a discrete choice framework to model how an individual chooses from a given set of alternatives, and analyzed three choice set optimization problems corresponding to agreement, dis- agreement, and promotion and under random utility models for discrete choice. We proved that the choice set optimization problems are NP-hard in general for these models, but we identify natural restrictions of the problems under which they become computationally tractable. These restrictions reveal a fundamental boundary: *promoting a particular decision within a group is easier than minimizing or maximizing conflict*. More specifically, we showed that restricting the choice models can make the promotion problem tractable while leaving agreement and disagreement NP-hard, indicating that the interaction between individuals introduces significant complexity to choice-set optimization. We also developed approximation algorithms for the NP-hard case and used this to analyze how effective one could be at influencing decisions using real-world data.

Extending this research, we are now considering context effects in social networks, using data about how the social network evolves over time (Tomlinson and Benson, In preparation, 2020). In this research, the "choice set" is a group of individuals with whom someone might communicate in a social network. We study context effects in terms of common social network features such as degree and reciprocity. We have a promising new "mixture model," where each mixture component corresponds to different groups of people who are swayed by different types of context effects.

Reasoning with neural networks

Another part of the project is understanding what types of reasoning can be done under computational constraints. One way of understanding computational constraints is to see what is feasible by a computer program and to understand the limitations of such computations. Neural networks pro- vide a computational framework that provides human-comparable (or better) performance in tasks such as speech recognition or object detection. Despite their success, a criticism of such techniques is that they are limited to low-level tasks as opposed to more sophisticated reasoning. This gap has drawn analogies to the difference in so-called "System 1" (i.e., low-level perception and intuitive knowledge) and "System 2" (i.e., reasoning, planning, and imagination) thinking from Kahneman et al. in social psychology. Proposals for moving towards System 2 reasoning in machine learning systems involve creating new abilities for composition and combinatorial generalization.

One approach for augmenting neural networks with these capabilities is performing relational reasoning over structured representations, such as sets. For relational reasoning, these systems are commonly split into two stages: (i) a perceptual stage that extracts structured sets of vector representations, intended to correspond to entities from the raw data, and (ii) a reasoning stage that uses these

representations. Typical differentiable methods directly map the input to latent features using a feedforward neural network and partition the latent features into a set representation for downstream reasoning. However, we showed that these methods have a fundamental flaw — at a high level, if there exists a continuous map that generates inputs from entities, then any function that can map such inputs to a list representation of the entities must be discontinuous. For relational reasoning tasks, this implies the perceptual stage must contain discontinuities. Using a continuous feedforward network to approximate discontinuities leads to representation issues (Fig. 1).

To address these issues, we introduced a *Set Refiner Network (SRN)*, a novel module that iteratively refines a proposed set using a set encode (Huang et al., submitted, 2020). The main idea of our module is that *instead of directly mapping an input embedding to a set, we instead find a set representation that can be encoded to the input embedding*. This procedure is a better model for discontinuities and can produce more meaningful set representations (Fig. 1). In extensive, we demonstrated that the SRN can effectively decompose entities in a controlled setting. Furthermore, we found that incorporating our SRN into state-of-the-art relational reasoning pipelines in computer vision, reinforcement learning, and natural language processing tasks improves prediction accuracy and robustness substantially.

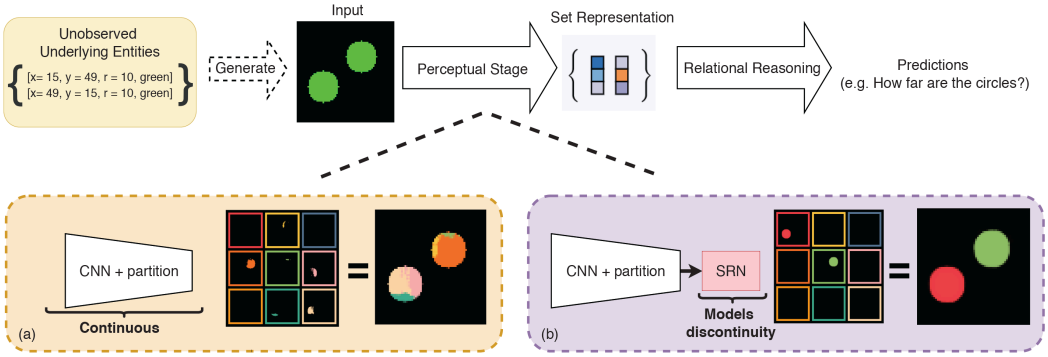


Figure 1: (Top row): Illustration of a standard relational reasoning paradigm. A perceptual stage that effectively recovers underlying set structure must be discontinuous. The second row shows a visualization of the perceptual stage for (a) existing methods and (b) our proposed Set Refiner Network. Each of the nine color-coded panels corresponds to one of nine set elements. As existing methods can only represent continuous functions, it is not able to recover the underlying object structure of the image. However, our approach can model discontinuities and thus can recover the

This research was conducted in collaboration with Facebook, and they are currently testing these ideas in their own machine learning pipelines. Published and accepted papers:

- Choice Set Optimization Under Discrete Choice Models of Group Decisions. Kiran Tomlinson and Austin R. Benson. To appear in *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.
  Code release: https://github.com/tomlinsonk/choice-set-opt.

- Neural Jump Stochastic Differential Equations. Junteng Jia and Austin R. Benson. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. Code release:
  https://github.com/000Justin000/torchdiffeq/tree/jj585.

- Frozen Binomials on the Web: Word Ordering and Language Conventions in Online Text. Katherine Van Koevering, Austin R. Benson and Jon Kleinberg. *Proceedings of the Web Conference (WWW)*, 2020.
  Code release: https://github.com/ktvank/Frozen-Binomials.

Working papers:

- Set-Structured Latent Representations. Qian Huang, Horace He, Abhay Singh, Yan Zhang, Ser-Nam Lim, and Austin R. Benson. *arXiv:2003.04448*, 2020. (Under review.)
  Code release: https://github.com/CUVL/SSLR.

- Context Effects in Social Network Formation. Kiran Tomlinson and Austin R. Benson. 2020. In preparation.

Biases in decision making by machines and effect of temporal ordering of words
We have been continuing our studies of bias in both human and algorithmic decision-making. A fascinating source of hidden language signals for some of these investigations has been in the orderings of words. There is inherent information captured in the order in which we write words in a list. The orderings of binomials --- lists of two words separated by `and' or `or' --- has been studied for more than a century. These binomials are common across many areas of speech, in both formal and informal text. In the last century, numerous explanations have been given to describe what order people use for these binomials, from differences in semantics to differences in phonology. These rules describe primarily `frozen' binomials that exist in exactly one ordering and have lacked large-scale trials to determine efficacy. Online text provides a unique opportunity to study these lists in the context of informal text at a very large scale. In our recent work, we expand the view of binomials to include a large-scale analysis of both frozen and non-frozen binomials in a quantitative way. Using this data, we then demonstrate that most previously proposed rules are ineffective at predicting binomial ordering. By tracking the order of these binomials across time and communities we are able to establish additional, unexplored dimensions central to these predictions.

Human and Machine Decision-makers: Skill gaps
Over the past year, we have also been studying gaps in skill between humans and AI systems, in domains that require high levels of expertise. Chess is a very useful model system for this, in that it is an extensively recorded domain in which humans of different

skill levels perform a set of cognitively demanding tasks; and AI systems have by now far surpassed the best human performance. A natural question is whether it is possible to "attenuate" these AI systems so as to match not just the strength of typical human chess players, but also their move-by-move behavior. We find that simply weakening a standard chess-playing algorithm, the traditional method for matching the playing strength of different levels of human players, does not produce systems that match the moves of human players. In contrast, we introduce a new system, a customized version of Alpha-Zero trained on human chess games, that predicts human moves at a much higher accuracy than existing engines, and can achieve maximum accuracy when predicting decisions made by players at a specific skill level in a tunable way. For a dual task of predicting whether a human will make a large mistake on the next move, we develop methods that significantly outperform competitive baselines. Taken together, our results suggest that there is substantial promise in designing artificial intelligence systems with human collaboration in mind by first accurately modeling granular human decision-making.

K. Van Koevering, A. Benson, J. Kleinberg. Frozen Binomials on the Web: Word Ordering and Language Conventions in Online Text. *Proceedings of 29th International World Wide Web Conference*, WWW 2020.

R. McIlroy-Young, S. Sen, J. Kleinberg, A. Anderson. Aligning Superhuman AI and Human Behavior: Chess as a Model System. *Proc. 24th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, 2020.

M. Raghavan, S. Barocas, J. Kleinberg, K. Levy. Mitigating Bias in Algorithmic Employment Screening: Evaluating Claims and Practices. *Proc. ACM Conference on Fairness, Accountability, and Transparency (FAT*),* 2020.

H. Heidari[1], J. Kleinberg. Optimally Allocating Opportunities in a Dynamic Model of Intergenerational Mobility: *Markov Decision Policies and Affirmative Action, submitted*

R. McIlroy-Young, S. Sen, J. Kleinberg, A. Anderson. *Learning Personalized Models of Human Behavior in Chess submitted for publication*

[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

[1] Jadbabaie's former student doing a Postdoc with Kleinberg and now joining CMU's Machine Learning department and Institute for Software research as faculty

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:

Group decision making with sampled beliefs: A cognitive model
Following the main thesis of our MURI project, we have studied resource-constrained group decision making as cognition in a Bayesian framework in which individuals can only sample from the belief distributions rather than fully computing it. We suspect that these very severe approximations are meta-rational in a certain sense: when the cost of computation is taken into account, they may reflect the ideal tradeoff of cost/speed and accuracy. By giving us insight into realistic cost functions for people, we can then apply the framework to other applications. Putting together the insights gained from considering specific scenarios such as those discussed above, our goal is to construct a foundation for a general theory of optimal resource-bounded behavior that would allow us to explain and predict human behavior in a wide variety of settings.

Recent literature in cognitive science (e.g. *by Gershman, Horvitz, & Tenenbaum,* 2015; *Vul, Goodman, Griffiths, &Tenenbaum, 2014* and the references therein as well as the work reported in the previous section by *MH Tessler and Tennenbaum 2020*), suggest that people often appear to make decisions based on just one or a few samples from the appropriate posterior probability distribution rather than using the full distribution. Although sampling-based approximations are a common way to implement Bayesian inference, the very limited number of samples often used by humans seem insufficient to approximate the re-quired probability distributions very accurately.

We have studied the consequences of agents communicating only *a single random sample* of their belief distribution, and ask: If people are making decisions based on a single random sample from their belief, under what conditions social learning is still possible? More specifically, motivated by the limited cognitive and communication resources available to decision-makers in large group settings, we investigate distributed learning and evolution of beliefs in a framework where individuals can only communicate samples from their beliefs (or a sparsified version of it), rather than the full distribution. Questions that immediately follow are: what are the conditions on sampling strategies for learning, non-learning, and mislearning to happen and whether/how they depend on the structure of the in-formation exchange network? Is learning even possible under such a scheme? Surprisingly, our preliminary findings suggest that, provided certain conditions are satisfied by private signaling likelihood functions, uniform sampling can result in mislearning (that is agents ending up assigning zero probability to the true state)with a positive probability, in the same setting where communicating full beliefs would result in learning the true state almost surely.

Furthermore, we have shown that when agent have enough self confidence in their own belief, even a single random sample from the posterior distribution of neighboring agents is enough to learn the optimal action.

While this update was in the case where agents have a recency bias and cannot recall the past, and are "pseudo-rational", we have also studied the fully Bayesian setting when the likelihoods and the uncertainty are all Gaussians. In this case, we have characterized completely the learning dynamics in the setting when agents are Bayesian, and the environment is Gaussian, demonstrating how the network structure can affect the speed of learning and the appearance of 'groupthink'. Moreover, in this setting, the updates are computationally efficient and can be carried out by agents in reasonable (i.e. polynomial) time in the size of the network.

We are currently working on determining what kind of samplings will result in learning, mislearning, group think, and polarization.

- Rabih Salhab, Amir Ajorlou, Josh Tenenbaum, and Ali Jadbabaie, "Social Learning with Sparse Belief Samples*," In Proceedings of 42nd Annual Virtual Meeting of the Cognitive Science Society*, *CogSci 2020,*
- Yash Deshpande and Elchanan Mossel, "Bayesian Social Learning from Samples in a Gaussian Environment," *in preparation*

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[2]. Bayesian Decision Making in Groups is Hard, *Operations Research, forthcoming in 2021*

- J. Hązła, A. Jadbabaie, E. Mossel, and M. A. Rahimian. Reasoning in Bayesian opinion exchange networks is PSPACE-hard. In A. Beygelzimer and D. Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1614–1648, Phoenix, USA, 25–28 Jun 2019. PMLR.

- E. Mossel, M. Mueller-Frank, A. Sly, and O. Tamuz. Social learning equilibria. *Econometrica*, 88(3):1235–1267, 2020

Probabilistic dependency graphs
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 and are more modular than Bayesian Networks (BNs), in that they make it easier to incorporate new information and restructure the representation. They show by example how PDGs are an especially natural modeling tool. They provide three semantics for PDGs, each of which can be derived from a scoring function (on joint distributions over the variables in the network) that can be viewed as representing a distribution's incompatibility with the PDG. For the PDG corresponding to a BN, this function is uniquely minimized by the distribution the BN represents, showing that PDG semantics extend BN semantics. A slight variant of the scoring function (which does not affect the score of a distribution that represents a BN) yields the variational free energy of general factor graphs; this can be used to explain how PDGs and factor graphs generalize BNs in orthogonal directions.

- Geffner and J. Y. Halpern, Lower bounds on implementing mediators in asynchronous systems, manuscript in submission, 2020.

[2] Jadbabaie's former student doing a postdoc with E. Mossel and now joining University of Pittsburgh as faculty

Group decisions, opinion exchange, persuasion, and role of disagreements
We have been continuing to study the dynamics of agreement and disagreement within groups, and have investigated network models for how adversaries can potentially inject disagreement into a group. We investigate the properties of such attacks, considering optimal strategies both for an adversary seeking to create disagreement and for potential countermeasures to make the group more robust against such attacks. Surprisingly, we show that for analyzing the effect of these attacks, the entire spectrum of the underlying interaction network is relevant for evaluating the adversary's power; this contrasts with related problems in which it is typically only the extreme eigenvalues that govern the behavior. We also consider the algorithmic task of a network defender to mitigate these sorts of adversarial attacks, and show connections between the geometry of this problem and methods in convex programming.

Furthermore, we have considered a popular scenario for distributed inference and learning whereby agents receive initial private (local) signals, and aim to take the best action given the aggregate observations of all other agents. Agents have access to estimates of their neighboring agents which are repeatedly updated. We study the implications of the information structure, the choice of the probability distributions for signal likelihoods and beliefs, and agents' recent memory. These factors determine the structure of the update rules that express the future actions of agents in terms of their observations. We analyze the resulting linear updating of actions based on the observed neighboring actions and show that many inefficiencies arise in group decisions due to repeated interactions between individuals leading to choice-shifts toward extreme actions. Nevertheless balanced, regular networks demonstrate a measure of efficiency in terms of aggregating the initial information of individuals.

- J. Gaitonde, J. Kleinberg, E. Tardos. Adversarial Perturbations of Opinion Dynamics in Networks. *Proc. 21st ACM Conference on Economics and Computation (EC), 2020*.

- M. Amin Rahimian, Amir Tohidi, and Ali Jadbabaie, " Bayesian Foundations of linear opinion dynamics," in submission

Opinion polarization
Following up on our work on persuasion and polarization and to better understand issues related to group decision making dynamics, we introduce a simple, geometric model of opinion polarization. Our model can be construed as a model of political persuasion, as well as marketing and advertising, utilizing social values. It focuses on the interplay between different topics and wide-reaching persuasion efforts in the media. We demonstrate that societal opinion polarization in our model often arises as an unintended byproduct of organizations attempting to sell an idea. We discuss some exploratory examples, analyzing how polarization occurs and evolves. We also examine some

computational aspects of choosing the most effective means of influencing agents in our model, and the connections of those strategic considerations with polarization.
More specifically, we consider two scenarios: either there is one entity (an influencer) trying to persuade agents to adopt their opinion or there are two competing influencers pushing different agendas. With respect to the timescale of interventions, we also consider two cases: the influencer(s) can apply arbitrarily many interventions, i.e., the asymptotic setting, or they need to maximize influence with a limited number of interventions, i.e., the short-term setting. The questions asked are: (i) What sequence of interventions should be applied to achieve the influencer's objective? (ii) What are the computational resources needed to compute this optimal sequence? (iii) What are the effects of applying the interventions on the population's opinion structure? We give partial answers to those questions. The gist of them is that in most cases, applying desired interventions increases the polarization of agents.

- Jan Hązła, Yan Jin, Elchanan Mossel, Govind Ramnarayan, "A Geometric Model of Opinion Polarization," *arXiv:1910.05274, February 2020*

Persuasion

On the empirical front, we investigate how much political messages (newspaper articles, advocacy campaigns, etc.) change public attitudes on politically and morally charged issues. Moreover, does a single message change different people's attitudes very differently, perhaps even in opposite directions? Randomized experiments measuring the effect of political messages often find minimal evidence of differences between demographic groups yet these groups may not be the appropriate dimensions on which to find heterogeneity in political persuasion. We have looked to theoretical frameworks such as Moral Foundations Theory which describe latent factors of psychology such as people's moral values, and investigate whether these factors predict differences in the persuasive effect of messages.

Persuasion, surprise and information cascades

Finally, to better understand persuasion and its role in information spreading and other collective phenomena, we study a model of online news dissemination on a Twitter-like social network. Agents with heterogeneous Gaussian priors would decide whether to share a piece of news/new information they receive with their followers. Each agent makes a single sharing decision based on whether the new information can persuade her followers to move their beliefs closer to hers in aggregate. The question is, will there be a sharing cascade? Will the new information go virtual? And what is the connection between the new information going viral and credibility/precision of the news.

At the micro-level, we demonstrate how *surprise* and *affirmation* motives naturally emerge from the utility-maximizing behavior of agents when persuasion is the main motive for sharing news. We characterize the dynamics of the news spread and derive necessary and sufficient conditions for emergence of a cascade, based on which we examine the precision level of new information that maximizes the likelihood of a

sharing cascade. In particular, we reveal the connection of the optimal precision[3] of news (i.e. how truthful the new informatio is) *to the diversity in perspectives* (variance of priors) and the *crowd wisdom* (i.e. whether aggregation of prior opinions concentrate on truth) for the well-connected social networks whose corresponding *line graph*[4] have an average degree of at least two. We show that in a well-connected social network, the diversity of perspectives facilitates the spread of news at all precision levels. In a highly-diverse population, a wide range of precision levels, including the truth, may trigger a cascade almost surely. When a population's prior views are moderately diverse or homogeneous, the level of surprise in the new information is a key variable in whether the new information causes a cascade In Particular, if the population is collectively unwise, i.e., there is a wide gap between the true state and the average of prior perspectives, the truth is sufficiently surprising to cause a cascade almost surely.

By contrast, a collectively wise population would filter out the truth if too homogeneous. Such a population, if moderately diverse, would single out the truth as the only news that will always go viral. By contrast, for the collectively-wise population, the truth would be filtered out if the perspectives are too homogeneous while the truth would spread as the only news that will always go viral if moderately diverse.}

Our results complement the empirical findings that support wider spread of inaccurate/false news compared to accurate information on social networks, providing a theoretical micro-foundation for utility-based news-sharing decisions. nWhile the above agents update their priors with Bayes rule, they do not interpret not receiving the news so in a sense they are not fully rational. How would fully rational agents behave in such a setting? To address this, 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 a 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. After observing the news, the source decides whether to disclose the news to subscribers. This decision is based on whether disclosure of the news persuades more subscribers to take the binary action the source favors. 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 such that they form a

---

[3] Precision of random variable is the reciprocal of its variance.

[4] The line graph of a graph is one in which nodes are edges and edges are formed when two edges are incident at a node in the original graph

subscription connection (the subscriber would follow the leader). Our results shed light on how individuals would 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. The theory may be applied to news markets in which readers subscribe to media outlets who may aim to influence the public opinions.

- Chin-Chia Hsu, Amir Ajorlou, and Ali Jadbabaie, "A Theory of Misinformation spreading in Social Networks," 2nd revision. To be submitted to *Review of Economic Studies (RESTUD), July 2020*
- Chin-Chia Hsu, Amir Ajorlou, Muhamet Yildiz, and Ali Jadbabaie, "Information Disclosure and Network Formation in News Subscription Services," *submitted to IEEE Conference on Decision and Control*

Thrust III: Modeling and Experimental Investigations

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.

The goal of this project is to introduce a new measure of habit strength and investigate the presence of habitual behavior in consumer choices. Various factors are in play in consumers' decision-making processes, one of which is the context of the decision that can trigger exiting habits. Habits are ubiquitous in human behavior. Studies show that between a third and a half of our daily decisions could be considered habitual. As a result, understanding the strength of habits, their formation and transformation, and how they are connected to other psychological processes is crucial to have an in-depth understanding of human behavior.

One of the main challenges is the definition and measurement of habits. Psychologists define them as "automatic processes that are learned from repeated responses, and are triggered through various contextual or mental cues". This definition revolves around the so-called habit loop that consists of three main elements: Contextual cues, repeated response, and the reward. In other words, people make certain decisions based on the context with a subsequent internal of external reward. The repetition of this process in the same context makes the whole loop more automatic and spontaneous . In accordance with this definition,

psychologists have proposed habit measures such as Frequency in context, and Self-Report Habit Index(SRHI).

In the Marketing and Economics literature, however, a different approach has been taken to study habits by measuring the effect of previous purchase decisions on current ones. Guadagni and Little add a consumer loyalty variable to their logit choice model which is the exponentially weighted sum of past purchases added to the other terms in the utility function. This term is also known as *consumer inertia or state dependenc*e. The inertia coefficient measures how more probable it is to buy the same product that one has purchased previously.

Nevertheless, shopping is a complex behavior and the so-called consumer inertia is affected by numerous mechanisms. To differentiate the habit element, we use store closures as an external shock that interrupts repeated behavior and can be used as a measure of habit strength. This idea is inspired by the habit discontinuity hypothesis which states that context changes in people's personal, social, or professional circumstances provide opportunities for conscious, planned behavior change. The key idea is that when a habit is blocked or suspended due to a change of context, the person may need and search for information or advice, and be open to alternative options.

To this end, we use Nilsen's store retail scanner and consumer panel data, containing detailed shopping information for around 600,000 American households and around 380,000 stores across the US since 2005. In order to identify closing stores, we compute the aggregate store weekly sales and single out the ones whose sales drop to zero at a certain time. We found 900 such stores during a 10-year period starting 2005. Then, by matching the resulting stores with the consumer panel data, we identify affected households who did a significant fraction of their shopping at those stores and regular visits until the closure. Our hypothesis was that since these households have been forced to experiment and learn in new shopping environments, they might on average have a decrease in their state dependence which could be a measure of shopping habits strength. By shopping habits, we refer to the extent that decision making is affected by the context of the regularly visited store.

We examined yogurt, as one of the most frequently purchased product categories. Then, we estimated a Mixed Logit choice model (a.k.a. Hierarchical Bayesian Model) with normal priors. These models are more flexible than Logit with fixed parameters since they allow for heterogeneity among panelists. As expected, not only we found a statistically significant decrease in state dependence comparing before and after the closure, but also very substantial compared with the average state dependence (more than a third of that). Note that the drop we are measuring here is a lower bar for the effect of closures because these households

didn't do all of their shopping in the closed stores. So only part of their behavior is affected by these events. Moreover, as a placebo test, we estimated the same model on a random set of households not affected by any store closure and did not find any statistically significant drop in their state dependence. This further reinforces our hypothesis that change of the context has a remarkable effect on households' behavior.

Finally, this framework can be used to measure the strength of habits whenever there is a serious context discontinuity that requires people to make new decisions. Particularly, the significant changes brought by the spread of COVID-19 can provide us with valuable data to test this model and gain better insight into the psychology of habits.

5. Awards, Honors, and Keynote talks by PIs (June 2019-June 2020)

- **Ali Jadbabaie** gave a remote talk at Rafael del Pinot in Spain.
- **Jon Kleinberg** is a recipient of a 2019 **Vannevar Bush Fellowship** from the Office of the *Undersecretary of Defense for Research and Engineering*
- **Elchanan Mossel** is a recipient of a 2020 **Vannevar Bush Fellowship** from the Office of the *Undersecretary of Defense for Research and Engineering*
- **Josh Tenenbaum** is a recipient of a 2019 Mac Arthur Fellowship from the *Mac Arthur Foundation*
- **Joe Halpern** was inducted into the *National Academy of Engineering*
- **Josh Tenenbaum** was elected to the *American Academy of Arts and Sciences*
- **Elchanan Mossel** delivered the 2nd Annual Peter Whittle lecture, Mathematics, University of Cambridge, Oct 2019.
- **Ali Jadbabaie** delivered a plenary talk at SIAM conference on Network Science in July 2020
- **Ali Jadbabaie** is scheduled to give a distinguished seminar at EPFL School of Engineering (postponed)
- **Ali Jadbabaie** was an Invited plenary Speaker at the Workshop on Social and Economic Networks: Polytecnico Torino, Turin, November 2019
- **Austin Benson** received a JP Morgan Chase & Co. AI Faculty Fellowship Award

- **Ali Jadbabaie** was an Invited Speaker at the Workshop on Deliberation, Belief Aggregation, and Epistemic Democracy, Neuville-sur-Oise, France, June 2019,

- **Ali Jadbabaie** was the co-organizer of the first and second Learning for Dynamics and Control Conference

6. Spending

Our project is on track to spend the entire increment that was delivered in early June. As of early June, we have spent 1.15M of the received 1.5M and we are on track to spend everything by the end of summer.

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.