REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
17-11-2020	Final Technical Report	09/2016 - 09/2020

4. TITLE AND SUBTITLE

Final Technical Report: A Comprehensive Research Content and Workflow Pipeline to Increase Openness, Reproducibility, and Prediction in Social Science Research.

5a. CONTRACT NUMBER
D17AC00002

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Alexander DeHaven, Timothy Errington, Ronald Brooks, Brian Nosek

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Center for Open Science, Inc.
210 Ridge McIntire Rd. Suite 500, Charlottesville, VA 22903

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Final Technical Report for the Center for Open Science's efforts on the Next Generation Social Sciences Program. COS assisted performing teams with making data open and workflows reproducible, evaluated the reproducibility of performers' efforts, and facilitated the preregistration of performer experiments and analyses.

15. SUBJECT TERMS

Reproducibility, Open Science, Social Behavioral Science, Preregistration

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			**19b. TELEPHONE NUMBER** *(include area code)*

Center for Open Science (COS)
Next Generation Social Sciences (NGS2) Program
Cooperative Agreement # D17AC00002
Final Technical Report

Date of Report: 11/17/2020

Project Title:	"A Comprehensive Research Content and Workflow Pipeline to Increase Openness, Reproducibility, and Prediction in Social Science Research"
Program Manager:	Lt. Col. Philip Root, DARPA Defense Sciences Office (DSO)

Submitted by:
Alex DeHaven, Project Manager
Center for Open Science, Inc.
210 Ridge McIntire Rd. Suite 500, Charlottesville, VA 22903
Email: alex.dehaven@cos.io

Distribution List and Email Addresses

Name	Title	Email Address
Philip Root	Deputy Director	philip.root@darpa.mil
Kristen Fuller	DSO Assistant Deputy, Program Management (ADPM)	Kristen.Fuller@darpa.mil
Lisa Troyer	Agreements Office Representative (AOR)	lisa.l.troyer.civ@mail.mil
Judy Williams	Cooperative Agreement Administrator	judy_williams@ibc.doi.gov
ONR Atlanta	Office of Naval Research	ONR_ATLANTA@navy.mil

Security Classification -- Unclassified

Cycle 1

Table 1. Cycle I Deliverables

Deliverable	Comment	Links
COS coordinates with DARPA and performers to kickoff NGS2 program	COS facilitated kickoff and storage and sharing of performer presentations	Kickoff Presentations: https://osf.io/6uzw2/
COS coordinates and manages initial training, pre-registration, experimental design for Cycle 1.	COS met with performers to provide preregistration and OSF training.	**UPenn Cycle I** Experiment 1: https://osf.io/5v3u6/?view_only=4df4b7bcf46a41b893bc63b9b499c08e Experiment 2: https://osf.io/rcs3u/?view_only=4df4b7bcf46a41b893bc63b9b499c08e Experiment 3: https://osf.io/bmfua/?view_only=4df4b7bcf46a41b893bc63b9b499c08e **Gallup WITNESS Cycle I** Experiment 1: https://osf.io/zksqj/ Experiment 2: https://osf.io/yvsj3/ **Berkeley Cycle I** https://github.com/Dallinger/ngs2-cycle1 **Virginia Tech Cycle I** https://osf.io/4mp5q/?view_only=32a95b89a98b42868a666dbc58ec855e
COS coordinates performers' initial experiments and reproducibility efforts.	COS met with performers and provided guidance on experimental design, focusing on reproducibility and transparency.	See attached "NGS2 Cycle 1 Reproducibility Round Robin.docx" for instructions. For summary results, see "NGS2_Cycle1_RoundRobin Results_Combined.docx"

COS presents results of Cycle 1 predictions feedback on Cycle 2 plans.	COS gathered Cycle I reproducibility efforts results and provided guidance for Cycle II.	Cycle 1 presentation of results and feedback: https://osf.io/fb576/

Table 2. Cycle I Tasks

Task	Comment
Receive information from DARPA on ETE and Enabler proposed experiments, initial hypotheses and academic disciplines.	COS provided feedback and guidance on proposed experiments, focusing on transparency and reproducibility.
COS receives information from ETE and Enablers on work plan and technical approach.	COS received information on planned work and approach, providing advice where needed. See preregistration links in Table 1 above.
COS advises performers on IRB documentation and protocol manuscripts.	COS provided advice focused on transparency and reproducibility.
COS provides testing and evaluation plan to DARPA, ETE and Enabler participants.	COS crafted testing and evaluation plan with input from participants. See Table 1 for further information.
COS confirms and/or expands academic journal partners, as needed, to manage a Registered Reports process.	COS attempted to implement the Registered Reports process, but was unable to due to timeline considerations. See below for additional details.
COS provides intensive OSF workflow and pre-registration training to all ETE and Enabler performers.	Training was provided to all performer teams. See Table 1 for links to performer preregistrations and projects.
COS ensures all ETE and Enabler performers complete pre-registration of their experimental design on the OSF.	COS worked with performers to provide feedback on preregistrations and ensured that all were stored on OSF. See Table 1 above for links.
Work with appropriate ETE/Enabler performers and journals to manage a Registered Report process.	COS attempted to implement this process, but it ultimately did not meet timeline requirements for the program. Instead, an adapted process was implemented.
COS assists performers with capturing experimental data and workflow in the OSF.	COS provided training on using OSF, including uploading data and other artefacts.

COS coordinates sharing of experimental design and data to other NGS2 performers.	COS served as a central hub for sharing between performers, passing designs between teams for feedback and review.
COS assists performers to reproduce another ETE team's experiment.	COS helped performers reproduce other teams' experiments to check for reproducibility.
COS supports performers to collect initial experiment and reproduced experiment results.	COS assisted performers in collecting data and generating results.
Assess any emerging OSF infrastructure development needs of ETE groups.	COS met regularly with performers to assess needs and develop collaborations or integrations as needed.
COS facilitates comparing experiment predictions with observed outcomes in both.	COS developed review protocol for comparing results to predictions. See ReproRubric information in Table 1.
COS prepares summary of Cycle 1 results for performers to help plan design for Cycle 2.	See link in Table 1 for Cycle 1 results and design plans for Cycle 2.

Cycle 2

Table 3. Cycle 2 Deliverables

Deliverable	Comment	Links
COS coordinates and manages, Cycle 2 pre-registration and initial experimental design	COS worked with performers to design and register experimental designs on OSF.	**Gallup Cycle 2** https://osf.io/vmk4b/ **Prediction Market Cycle 2** https://osf.io/8p2ac **Berkeley Cycle 2** https://osf.io/be67q/ **UPenn Cycle 2** See attached zipped folder for UPenn preregistration and reproducibility package "Penn Cycle 2 Reproducibility Round Robin.zip"
COS coordinates performers Cycle 2 experiments and reproducibility efforts.	COS worked with performers to ensure experiment design and results were stored on OSF. COS assessed the reproducibility of performer experiments.	See spider graphs in "NGS2 Cycle 2 ReproRubric Radar.xlsx"
COS presents results of Cycle 2 predictions, feedback, and updated plans for Phase 2.	COS gathered results of Cycle 2 efforts and presented them to performers.	Cycle 2 presentation (https://osf.io/r3a6n/)

Table 4. Cycle 2 Tasks

Task	Comment
COS provides OSF workflow Training/Pre-registration to all Enabler performers and ETE performers who wish to review the OSF capabilities.	COS continued to provide training to performer teams.
COS advises performers on IRB documentation and protocol manuscripts.	COS provided advice focused on transparency and reproducibility.

COS coordinates with all ETE and Enabler performers complete pre-registration of their experimental design.	COS worked with performers to ensure their experimental designs were adequately defined and registered on OSF.
COS works with appropriate ETE/Enabler performers and journals to manage a Registered Report process.	COS attempted to implement the Registered Report process, but an adapted system was devised instead to meet timeline constraints.
COS assists performers with capturing updated experimental design, workflow and experimental data in the OSF.	COS continued to provide training on using OSF, including uploading data and other artefacts.
COS coordinates sharing of Cycle 2 experimental design and data to other NGS2 participants.	COS served as a central hub for performer teams to share data between teams.
COS assists participants in reproducing another ETE team's Cycle 2 experiment.	COS gathered performer protocols and passed them between groups while assisting in attempting to reproduce them.
COS will work with all performers to collect initial experiment and reproduced experiment results.	COS assisted performers in collecting data and generating results.
Iterate work done in Cycle 1 and Cycle 2 (compare initial experiment predictions, experiment outcomes and reproducibility).	COS performed comparisons between prediction and outcome for performers.
Assess technology needs of ETE groups for Phase 2.	COS met with performers to assess needs and develop integrations where possible.
Present results of Cycle2 predictions and update plans for Phase 2.	COS presented Cycle 2 results and plans for Phase 2. See Table 3 for presentation link.
Update Technology, testing and evaluation plan for Phase 2.	COS updated testing and evaluation plan for Phase 2.

Cycle 3

Table 5. Cycle 3 Deliverables

Deliverable	Comment	Links
COS coordinates and manages, Cycle 3 pre-registration, experimental design, execution and experiment reproduction.	COS continued to work with remaining performers to preregister study designs and analysis plans.	**Gallup Cycle 3** Polycraft World: https://osf.io/q7v4x/ Boomtown: https://osf.io/wbckq/ **Berkeley Cycle 3** https://osf.io/q2m9h?view_only=e48ceb31704549dba78ef532b2716df5
COS presents Cycle 3 experimental results to DARPA and performers.	COS gathered Cycle 3 results and presented them to performer teams.	Cycle 3: https://osf.io/ak7qj/)

Table 6. Cycle 3 Tasks

Task	Comment
COS advises performers on IRB documentation and protocol manuscripts.	COS provided advice focused on transparency and reproducibility.
COS provides OSF workflow Training/pre-registration review available to performers.	COS continued to provide training to performer teams.
COS ensures all ETE performers complete pre-registration of their experimental design.	COS worked with performers to ensure their experimental designs were adequately defined and registered on OSF.
COS supports ETE performers and journals to manage a Registered Report process.	COS attempted to implement the Registered Report process, but an adapted system was devised instead to meet timeline constraints.
COS assists performers with capturing experimental data and workflow in the OSF.	COS continued to provide training on using OSF, including uploading data and other artefacts.
COS coordinates sharing of experimental design and data to other NGS2 performers.	COS served as a central hub for performer teams to share data between teams.

COS assists performers in reproducing another ETE team's experiment.	COS gathered performer protocols and passed them between groups while assisting in attempting to reproduce them.
COS will work with all performers to collect initial experiment and reproduced experiment results.	COS worked with performers to collect original results and reproduced results for comparison.
Iterate work done in Cycle 1, 2 and 3.	COS continued to iterate on previous work.
Assess technology needs of ETE groups.	COS met with performers to assess needs and develop integrations where possible.
Update Technology, testing and evaluation plan.	COS reviewed testing and evaluation plan.
Present results of Cycle 3 predictions and outcomes.	COS presented Cycle 3 results. See Table 5 for presentation link.

Cycle 4

Table 7. Cycle 4 Deliverables

Deliverable	Comment	Links
COS coordinates and manages, Cycle 4 pre-registration, experimental design, execution and experiment reproduction.	COS continued to work with remaining performers to preregister study designs and analysis plans.	**Gallup Cycle 4** https://osf.io/cpq3a/ **Volunteer Science Cycle 4** https://osf.io/3en6j
COS presents final results NGS2 program	COS gathered available Cycle 4 results to present to performers. Due to delays, Cycle 4 was not completed.	Final Presentation: https://osf.io/ak7gj/

Table 8. Cycle 4 Tasks

Task	Comment
COS presents any updates on the testing and evaluation process to performers.	COS worked with performers on evaluation process. Due to the shortened timeline, this was primarily done with the Gallup team.
COS provides updated OSF workflow training/pre-registration review to performers as needed.	COS worked with Gallup and Volunteer Science to preregister their predictions on OSF. See Table 7 above for links.
COS ensures all ETE performers complete pre-registration of their experimental design.	COS worked with Gallup and Volunteer Science teams to register designs. The Berkeley team pivoted in Cycle 4 and did not register on OSF.
COS will work with appropriate ETE performers and journals to manage a Registered Report process.	The Registered Report process was not suitable for the NGS2 program, so this was not attempted.
COS assists performers with capturing experimental data and workflow in the OSF.	COS worked with the Gallup team to capture experimental data on OSF.
COS coordinates sharing of experimental design and data to other NGS2 performers.	COS worked with Gallup team to share designs with remaining performer teams.
COS assists performers in reproducing another ETE team's experiment.	Due to changes to performer foci and remaining time, this was not pursued.

COS will collect and evaluate Cycle 4 experiment and reproducibility results.	Due to limited time remaining on the program, this was able to be completed.
Review, work done in Cycles 1,2,3, and 4.	Cycles 1-4 were reviewed and general lessons learned were presented to teams.
Provide DARPA and all performers with NGS2 results.	Results of Cycle 3 were provided, but Cycle 4 was not able to be provided due to remaining time.
Document use of technology, research practices, testing and evaluation plans and changes that occurred in NGS2.	This documentation was done throughout the program and can be seen in the deliverable links in the above tables.
As directed by DARPA, assist in the transition of products to the wider social science research community and DoD.	COS is currently in discussions with Charles Rivers Analytics to investigate potential extensions and transitions.

Other Efforts

Technical Exchange on Complex Social Systems (TECSS)

Table 9. TECSS Deliverables

Deliverable	Comment	Links
List of confirmed planning group members.	COS gathered a list of diverse experts to participate in the TECSS conference.	See "TECSS Project Overview.pdf" for list of participants
Complete literature review.	COS staff completed a literature review ahead of the TECSS conference.	Literature Review: https://osf.io/thbgd/
Complete set of presentations, discussion notes, and other materials from the technical exchange preserved on OSF.	All presentations, notes, and other generated materials were stored on TECSS OSF project pages	TECSS OSF Page: https://osf.io/re5vt/
Papers covering all focus group topics posted to OSF Preprints.	Each of the 5 Working Groups were charged with crafting a manuscript capturing the group's topic and discussions.	Group1 : https://osf.io/u6vz5/ - In review at Harvard Data Science Review Group 2: https://osf.io/5md6u/ Group 3: https://psyarxiv.com/j8b9a - In review at Harvard Data Science Review Group 4: https://osf.io/932qp/ Group 5: https://osf.io/preprints/socarxiv/vncwe/ - In review at Harvard Data Science Review

Table 10. TECSS Tasks

Task	Comment
Establish a Planning Group of five thought leaders in various social sciences by April 13, 2018 to plan details of the technical exchange and inform the structure of the literature review for use as background material.	Planning group was recruited and the meeting was planned before the April 13 deadline. This group also helped structure the literature review that was conducted by Courtney Soderberg at COS.
Prepare a literature review by June 18, 2018 to describe the current understanding of the landscape and limits of theory, methodology, statistics and inference, aggregating evidence, and metascience for investigating complex social systems.	The literature review was completed in time by COS and shared with invited participants.
Host a 2-day technical exchange in Charlottesville, VA with leading experts from across the social sciences and beyond, where participants will divide into five focus groups based on the five topics in Task 1.2 and consider multiple dimensions for each topic including description, prediction, explanation, and control; level of analysis (individuals, groups, systems, cultures); practical versus theoretical limits; and ethical limits.	The TECSS meeting was hosted at COS in Charlottesville, VA. The meeting took place August 28-29, 2018.
Manage focus group efforts to author papers to translate the consideration of theoretical and practical limits of research on complex social systems into specific opportunities and potential high-impact solutions for advancing the state of the art for each topic.	Each group generated a working paper that was developed into a complete manuscript on the assigned topics.

Model-Based Research and Reproducibility Workshop (MBR2)

Table 11. MBR2 Deliverables

Deliverable	Comment	Links
Complete set of any presentations, discussion notes, and other materials from the workshop preserved on OSF.	All notes and other generated materials were stored on MBR2 OSF project pages	**MBR2 OSF Page** https://osf.io/bwdsr/ **Group 1** https://osf.io/jmy9c/ **Group 2** https://osf.io/vsy8u/ **Group 3** Preregistration template https://osf.io/cte9k/ Notes https://osf.io/6nvfp/
Content Development for Open Science Knowledge Base (OSKB)	OSKB Development Team collaborated with OER Commons to design and launch the OSKB knowledge base interface. Actions included establishing collections and groups to organize OSKB content. Bulk uploads of identified relevant content was added from the foundation of the knowledge base. OSKB Development Team wrote unique resources to support contributors to the knowledge base including a detailed flowchart for submitting a resource and a resource eligibility checklist.	View the full OSKB and contributor guides here: https://www.oercommons.org/hubs/OSKB

Table 12. MBR2 Tasks

Task	Comment
Host Meeting	Meeting was hosted at COS in Charlottesville, VA. The meeting took place February 4-5, 2020.
OSKB Content Development	OSKB Development Team worked with OER Commons to design and launch interface. See Table 11 for more details.
Registration Workflow Alpha	A preregistration template and workflow was designed for application of mathematical and computational models. See Table 11 for link.

Areas for Improvement

ReproRubric

The Reproducibility Rubric (ReproRubric) was designed to serve as a grading rubric for a project's reproducibility across a number of facets. This rubric worked well initially, and its use resulted in clearer performer specifications of experimental designs. However, as the template saw further use and refinement, it became clear that additional work was needed to fully define each element of the rubric across the different facets. The work has tremendous future potential, but achieving this potential would require considerable buy-in from stakeholders. In the NGS2 program, this buy-in was attempted by engaging the performer teams, soliciting feedback on the mechanism by which their work and outputs would be evaluated. Unfortunately, even in these circumstances, adequate clarity and specificity was not achievable.

Registered Reports for DARPA

Registered Reports are an innovative article publication workflow designed to combat the positive publication bias seen across most academic journals. DARPA wished to utilize this workflow, which shifts the focus of peer review away from the results, to the experimental designs, proposed hypotheses, and analytical plans or frameworks. Ultimately, this workflow was not feasible for a program like DARPA where it is important to move, fail, and adjust fairly rapidly. The present timeline of Registered Reports did not mesh well with this research mode, as review would take considerable time, and during review no work could be done without it also needing to be sent out for review. Additionally, given the cutting-edge nature of the program's work, it was exceedingly difficult to find reviewers for the process that were both knowledgeable of the RR process and the subject matter, all while not having a conflict of interest.

In the end, an adapted Registered Reports workflow was invented. Performers would preregister their designs and proposed analyses, and these proposals would be passed between the different groups for review. This allowed some review of the plans by the performer groups without the timeline constraints or conflict of interest concerns. This did miss out on one key advantage of the Registered Report model: the in-principle acceptance feature. Since this was done without any academic journal's involvement, the manuscripts could not receive a promise to publish the results as long as the peer reviewed plans were adhered to. Luckily, these articles could be posted as preprints on OSF preprints.

In future DARPA programs, it would be advantageous to identify journals that could handle a much faster review process for conducting Registered Reports, or design an internal system as devised in NGS2. Programs that wish to utilize this model might also benefit by incorporating Registered Reports into the design of the program, so the timeline and milestones of the program are aligned with the journal(s) peer review process.

PROJECT OVERVIEW

The purpose of the Technical Exchange on Complex Social Systems (TECSS) is to:

1. Identify current or historical vs. fundamental limitations – practical, theoretical, technological, ethical, and others – to understanding and predicting complex social systems, and quantitatively and qualitatively characterize the source(s) of those limitations.
2. Describe practical, but high-risk and high-reward opportunities to manage, reduce, or eliminate limitations and develop potential high-impact solutions.

These goals will be achieved through the following activities:

1. Preparing literature review to provide background materials on the current understanding of the landscape and limits of social science research by focusing on 5 main questions

2. Hosting a two-day technical exchange in Charlottesville, Virginia, USA on August 28th and 29th, 2018 where participants will divide into five focus groups based on the five questions discussed in the literature review and consider multiple dimensions for each topic including:

- Description, prediction, explanation, and control such that optimal strategies and policies can be adopted
- Level of analysis (individuals, groups, systems, cultures)
- Practical versus theoretical limits
- Ethical limits

Discussions at the exchange will focus on the theoretical and practical limits of each dimension to identify high-impact opportunities for each topic.

3. Focus Groups will author papers, following the Technical Exchange, to translate theoretical and practical limits of research on complex social systems into specific opportunities and solutions. These papers should seek to answer the Heilmeier Catechism and will include descriptions of the necessary technological investments, risks of the approach, and overall magnitude of impact of each approach towards the upper limit of what is practically achievable. Papers will first be posted as preprints to solicit feedback from the broader community. Final papers will then be submitted for publication, potentially as a special issue or series.

FOCUS GROUPS

Group 1: What would be a process to systematically and semi-automatically quantify the features of complex social systems, which define the fundamental limits and intrinsic uncertainty to making predictions? What currently intractable problems could success here help address? What "Research Challenges" might be useful for motivating and sponsoring significant leap-ahead capabilities in this area?

Lead: Duncan Watts (Microsoft, Sociology)
Members:
- Emorie Beck (Washington University - St. louis, Personality Psychology)
- Elisa Bienenstock (Arizona State University, Mathematical Sociology)
- Jake Bowers (University of Illinois at Urbana–Champaign, Political Science, Statistics)
- Aaron Frank (RAND, Computational Social Science)
- Tony Grubesic (Arizona State University, Geography)
- Jake Hofman (Microsoft, Computational Social Science)
- Julia Rohrer (Leipzig University, Psychology)
- Matt Salganik (Princeton, Sociology)

Group 2: Could validated, objective evaluation criteria for social science research be used to increase the confidence of decision makers in using SB research to inform decisions? What might these criteria look like and how might they be validated? How would acceptance of these criteria impact the self-correcting nature of science? What currently intractable problems could success here help address? What "Research Challenges" might be useful for motivating and sponsoring significant leap-ahead capabilities in this area?

Lead: Simine Vazire (UC Davis)
Members:
- Alexander Etz (UC Irvine, Statistics)
- Robert Groves (Georgetown, Sociology, Statistics)
- Gary Klein (MITRE, Social Psychology)
- Rick Lempert (University of Michigan, Political Science, Criminology)
- Michelle Meyer (Geisinger, Research Ethics)
- Amy O'Hara (Stanford, Demography, Economics)
- Nancy Potok (U.S. Office of Management and Budget, Statistics)

Group 3: How can the magnitude of impact on research results by contextual factors be quantified in different social science disciplines or subdomains in order to predict the impact of contextual factors for a poorly defined or rapidly changing research space? What currently intractable problems could success here help address? What "Research Challenges" might be useful for motivating and sponsoring significant leap-ahead capabilities in this area?

Lead: Danny Goroff (Sloan Foundation, Mathematics, Economics)

Members:

Cristina Bicchieri (University of Pennsylvania, Philosophy, Psychology)
Kate Coronges (Northeastern, Sociology, Networks)
Neil Lewis (Cornell University, Psychology)
Anne Scheel (Technische Universiteit Eindhoven, Psychology, Methods)
Laura Scherer (University of Missouri, Psychology)
Gillian Tett (The Financial Times, Anthropology)
Josh Tucker (NYU, Politics)

Group 4: What combination of adaptable meta-analytical tools and techniques could be developed to measure progress across a population of related studies while also dealing with social, political and technological evolution? (Read: assessing current research and defining a set of tools and techniques that are adaptable over time). What currently intractable problems could success here help address? What "Research Challenges" might be useful for motivating and sponsoring significant leap-ahead capabilities in this area?

Lead: Joanna Korman (US Naval Research Lab, Experimental Social Psychology)
Members:

George Banks (UNC-Charlotte, Management)
Fiona Fidler (University of Melbourne, Philosophy)
Dan Foy (Gallup, Anthropology)
Boyoung Kim (Brown University, Social Psychology)
Matt Makel (Duke, Education, Methods)
Phil Schrodt (Parus Analytics, Political Science, Statistics)
Stuti Thapa Magar (Purdue, Industrial/Organizational Psychology)

Group 5: To what extent can SBS be "automated" to radically accelerate discovery and validation in understanding and predicting complex social systems? What dimensions of current work flows could be most impacted by Machine Learning (ML) /Artificial Intelligence (AI)? What currently intractable problems could success here help address? What "Research Challenges" might be useful for motivating and sponsoring significant leap-ahead capabilities in this area?

Lead: Tal Yarkoni (UT Austin, Psychology, Psychoinformatics)
Members:

Dean Eckles (MIT, Communication, Statistics)
James Heathers (Northeastern, Psychology)
Julia Lane (NYU, Economics, Public Policy)
Maggie Levenstein (University of Michigan, Economics)
Cosma Shalizi (Carnegie Mellon, Statistics)
Paul Smaldino (UC Merced, Cognitive & Information Sciences)
John Myles White (Facebook, Statistics, Computer Science)

TECHNICAL EXCHANGE AGENDA

Day 1: Tuesday, August 28th, 2018

0730 Registration/Breakfast - *Center for Open Science*
0830 Opening Remarks - Adam Russell, DARPA Program Manager
0900 Break out into Focus Groups
1230 Lunch
1330 Focus Groups continued discussions
1530 Coffee Break
1600 Discuss Day 2 action items
1630 Break for the day (Dinner on your own)

Day 2: Wednesday, August 29th, 2018

0800 Breakfast - *Center for Open Science*
0830 Focus Groups prep for presentations
0930 Beverage Break
0945 Focus groups presentations
1230 Working Lunch - incorporating feedback and developing an action plan/next steps for writing papers
1400 Closing Remarks - Brian Nosek, Executive Director, COS
1430 Wrap up
1530 Exfil

Please create a copy of this document, re-title it with your team name, and share it with brandon@cos.io This document contains view-only links to unpublished data, so please be cautious about sharing these links. Thanks!

For each team, follow the provided instructions and score the reproducibility of each analysis on a scale of 1-10 based on the following criteria:
10 - automatically reproducible
7-9 - reproducible following a documentation file, no running errors or missing information
5-7 - reproducible with minor issues (e.g. missing packages)
3-5 - reproducible with major issues (requires extensive communication with the code owner)
<3 - not reproducible

A few notes:

- Presumably obvious, but skip your own experiments.

- As mentioned at the PI meeting, if the code executes without errors then just spot check 5-10% of figures, text, and merged or cleaned up data against the original analyses provided.

- We understand that scoring within each range will be somewhat subjective based on your expectations, knowledge of the software, etc. but these scores will just be used as anonymized feedback to help each other do better in Cycle 2. If it doesn't work out well, then we'll adjust the strategy going forward.

UPenn

View-only link to project files
Data analysis instructions and Matlab instructions

Scores
Experiment 1 original report
Reproducibility score:

Experiment 2 original report
Reproducibility score:

Experiment 3 original report
Reproducibility score:

Errors and other feedback:

Gallup

2 methods are provided - R markdown and virtual machine. Choose 1 or both as you like, but we recommend knitting the R markdown file.

Scores
Experiment 1 instructions and original report (need to download)
Reproducibility score:

Experiment 2 instructions and original report (need to download)
Reproducibility score:

Errors and other feedback:

Virginia Tech

All data and code is packaged in this .zip file (need to download). A README is provided to load the reproducibility package, and the supporting tutorial.tex document is available to discuss the results and running the scripts. Original analysis figures are also provided for comparison.

Score
Reproducibility score:

Errors and other feedback:

UC Berkeley

Score
Part 1, DIFI validation. Launch the binder by clicking the launch button at https://github.com/Dallinger/difi-validation, or here:

Reproducibility score:

Part 2, Cycle 1 predictions and results. Launch the binder by clicking the launch button at https://github.com/Dallinger/ngs2-cycle1, or here:

Reproducibility score:

Errors and other feedback:

Prediction Scoring

Task 2: Procedure for prediction scoring (performed by ETE teams on their own project)

- ❑ Decide on appropriate statistical tests for comparing predictions to results
  - ❑ Prediction and result are both values: absolute numerical difference
  - ❑ Prediction and result are both distributions: Kullback–Leibler divergence
  - ❑ Prediction is a distribution and result is a value: likelihood / deviance
  - ❑ Prediction is a time-series of subjects' actions derived from a mathematical model: test of model adequacy following ML parameter inference
- ❑ Preregister the analysis plan using Open-Ended registration (no review)
- ❑ Perform the analysis and post a summary of the results to OSF

(copied from: https://docs.google.com/document/d/1izSEk6hlRSjXEYp3qpicQGx8kK-sKdUUh8dZjE-Zs4/edit?usp=sharing)

Note: As discussed at the PI meeting, there are different ideas about how to best score predictions. If you think your idea is better suited than the ones listed above, go with your idea. For the purpose of closing Cycle 1, we should not debate this heavily until we have some informative data (i.e. the results of Cycle 1 prediction scoring from each team).

Describe your prediction scoring results here, or provide a link to the summary posted on OSF:

Berkeley team

Part 1, DIFI validation
10, 10, 10 (**avg 10**)

Part 2, Cycle 1 predictions and results
2, 9, 5 (**avg 5.3**)
Comments:

- The `preprocess` notebook errored with `BadZipfile: File is not a zip file`. After trying a number of things, including looking at the master code in the Dallinger repo, I couldn't get data to load or run.
- For Part 2 Cycle 1 a message displayed: *Repository Dallinger/ngs2-cycle1 is taking a long time to load! See the logs for details*. The interactive visualization like Part 1 doesn't appear.
- The launch button did not generate a binder. There were no instructions. The individual Jupyter notebooks worked.

General comments

- For the results binders, it seems the analysis notebooks used mostly hard-coded data, which feels counter-intuitive for reproducibility. It should rely on the processed data, not just that the data can be processed.
- Neat design and self-explanatory workflow demonstrated via Jupyter notebooks.

Gallup team

Experiment 1
9, 6, 7 (**avg 7.3**)

Experiment 2
9, 7, 7 (**avg 7.7**)

General Comments

- No errors, the instructions were detailed and I had full replicability very quickly. Perhaps including OSF items in a .zip file like the Virginia Tech team did would allow for quicker download.
- Reproducibility information should contain the R packages needed for Experiment 1. Reproducibility affected by specific versions of packages. Would prefer open-source software implementation.
- Using R studio did not work unless package "pacman" was already installed. Perhaps, this is an error in the code that checks for and installs the package. Access to several of the links in Rmarkdown to OSF is forbidden. It otherwise worked.
- Local strata: Where is the strata docker image? I wasn't able to complete this task without it. The public strata work.

Penn team

Experiment 1
7, 8, 6 (**avg 7**)
Comments:
R script

- Executed with minor directory touch-ups

Stata do-file

- Whole group graphic: Stata output doesn't seem to match the report graphic
- Own team graphic: Stata output nearly matches the report graphic
- Other team graphic: Stata output is off in some respects from report graphic (bar positive when it should be negative)
- Unclear from the Stata-generated graphic alone which graphic/panel should be used; somewhat more clear by reading notes in the do-file which graphic is relevant, but still unclear which panel is important -- the 'P' or 'Y' panel.

Experiment 2
8, 2, 7 (**avg 5.7**)
Comments:
R script

- Both scripts executed with no problems

Stata do-file

- Executed with no problems; graphs are fine

Matlab script(s)

- No explicit instructions to run Matlab scripts.

Experiment 3
8, 8, 8 (**avg 8**)
Comments:
R script

- Both scripts executed with no problems

Stata do-file

- Graph 1 matches
- Graph 2 has the right data, but based on the Stata graphic and the report graphic, the 'other team' and 'whole group' values are swapped.

Matlab script(s)

- Instructions indicate that the second R script creates input for Matlab, but there are no instructions on what to do in Matlab; however, there are no additional graphics in the experiment 3 report, so this probably isn't a big deal.

General Comments

- As a user that isn't super familiar with R, I had some trouble figuring out what to do in the beginning of the scripts. The comments in the scripts were used in combination with the documentation, but perhaps including the script comments in the documentation might be a bit more helpful so I know which script comments are relevant to me. Perhaps including OSF items in a .zip file like the Virginia Tech team did would allow for quicker download *(note from Brandon: the slow download may have been a WaterButler issue on the OSF side)*

- There was some minor need to tweak the R scripts, but those are minor and easy to do
- The Stata graphics -- across all three experiments -- were poorly labeled and were not the same form as in the reports; this made for some guesswork on what graphics were relevant and what labels in the Stata graphics meant. Not a reproducibility issue per se, but certainly something to think about -- the graphics themselves could not 'stand on their own' in the Stata output as they could in the report output
- The use of Stata is a serious drawback. It's closed-source and working around that is a dealbreaker if one doesn't have access to Stata. I did, so it was manageable, but everything that was done in Stata is easily replicable in something like R, which as open-source software, means that anyone could pick up and run these analyses
- The use of Matlab likewise is a serious drawback. While there might have been some specific reasons to use Matlab (I don't know Matlab code incredibly well...) it should have been made compatable with Octave, an open-source implementation of Matlab. The `histcounts` function in Matlab is most closely implemented as `histc` in Octave, so the Matlab scripts alone could not run. Changing the function name, however, lead to parameterization changes, which I did not implement.
- Note, viewing documents online through OSF seriously distorts the visuals; none of the error bars are visible and certain glyphs are missing altogether; these documents needed to be downloaded for accurate rendering on a Mac, at least.
- Reproducibility is spread across multiple software, some of which are not open-source. Therefore limits the reproducibility issue. Makes assumptions of working knowledge for some software on the part of users. Would prefer a single software implementation and a script for binding different execution steps.

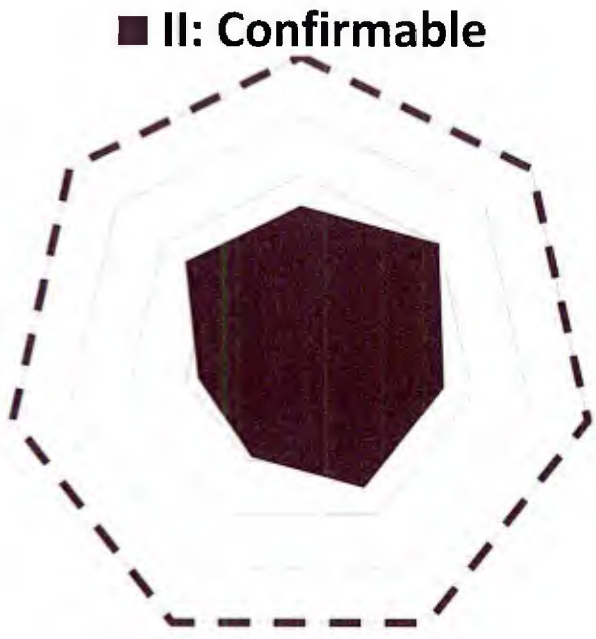
Virginia Tech team

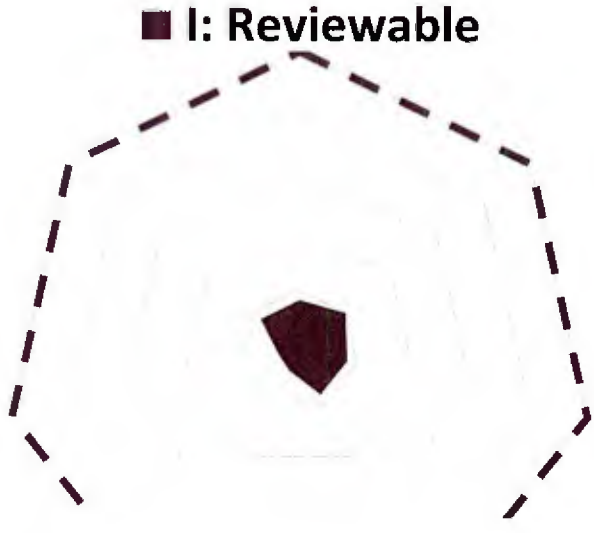
Experiment 1
8, 7, 9 (**avg 8**)

General comments

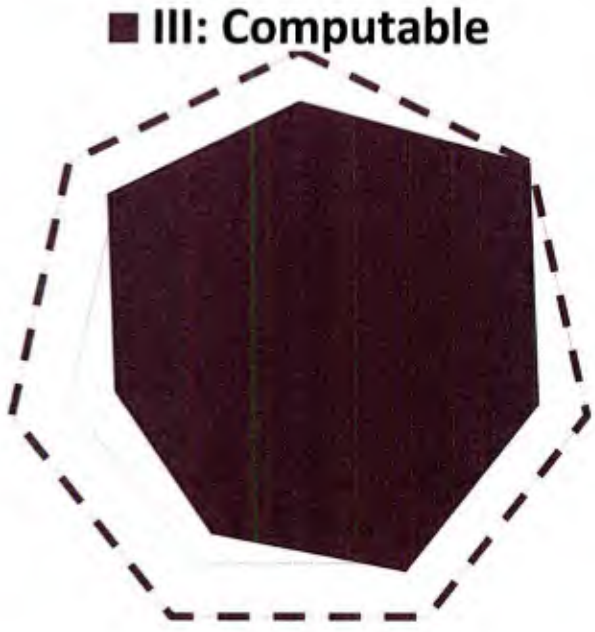
- No issues with reproduction, but still does not meet the "automatically reproducible" mark.
- Good idea providing a zip file in the OSF repo. However, the README instructions within the directory could provide a bit more specific information on what commands to run, my familiarity with virtual environments allowed me to complete the tasks, but perhaps this may have been difficult for other users. Aside from that, the run.all script is useful for getting everything setup. I initially had some issues with numpy, but was able to resolve that with some troubleshooting.
- Everything works following the documentation. I examined ~15% of the images I generated and they matched those supplied. The instructions were easily followed.

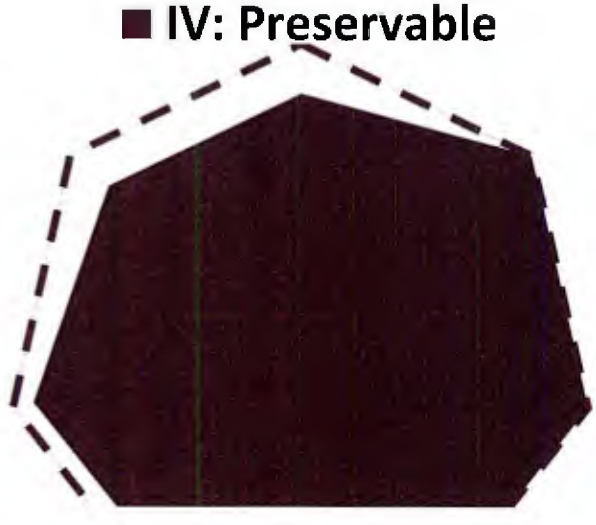
	Analysis Plan	Data Collection Plan	Data Collection	Data	Data Cleaning
I: Reviewable	0.166666667	0.2	0.166666667	0.166667	0.076923077
II: Confirmable	0.5	0.6	0.5	0.5	0.384615385
III: Computable	0.833333333	1	0.833333333	0.833333	0.692307692
IV: Preservable	0.833333333	1	1	1	0.923076923
V: Automatical	1	1	1	1	1
Cycle 1 target	1	1	1	1	1





Data Analysis	Workflow/ Reporting
0.071428571	0.166666667
0.357142857	0.5
0.642857143	0.833333333
0.928571429	0.833333333
1	1
1	1





- **V: Automatically verifiable**

Penn Team: NGS2 Cycle 2 Results and Reproducibility Round Robin Report

Innovation in an adversarial collective sensing with bots

0. Background

Individuals, groups, and organizations innovate, finding new and effective ways to communicate, plan, discover, construct, and play. These acts of innovation are the foundation of cumulative culture, by which successive generations build on achievements of the past, generating tools, techniques, and procedures of a sophistication that no individual or organization could have developed in isolation [12]. But not all individuals, organizations, and cultures are equally innovative, and not all time periods experience a steady accumulation of innovations. A rigorous understanding of what factors tend to promote innovation in groups, especially in competitive contexts, would have practical implications across a broad range of human endeavors.

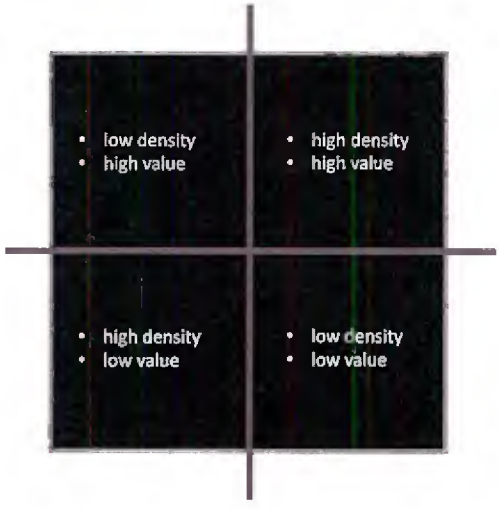
To bring the study of competitive innovation under experimental control, we developed with the Berkeley team an adversarial collective sensing games, which pit teams against each other in a competition to discover and extract resources from a shared environment. Adversarial collective sensing games are an elaboration of an experimental paradigm developed in the collective animal behavior [5] and cognitive science [28] literatures. In a collective sensing game, each position in a playing area is associated with a resource density, and any individual at that location observes a stochastic signal of the density at that point. The problem of collective sensing is to interpret these signals to discover resource-rich regions of the environment. Resource distributions can be complex spatially and temporally varying functions. Individuals view the locations of others in the environment, or others nearby, and may have some mechanisms for communication. Through observing each others' locations, players can learn from each other, or players can hinder each other by using location and communication in strategic adversarial capacities.

Whereas the Berkeley NGS2 team tested hypotheses about innovation in the adversarial collective sensing with human participants, the Penn NGS2 team tested hypotheses in the same basic game in the presence of "bots" whose programming determine the experimental condition.

1. Experimental Setup

- We ran experiments each with a single human and 99 bots.
- Players were separated into two equal sized "teams" (yellow and purple)
- Resources were heterogeneously distributed in time and space. There were 4 patches (resources re-occur in time within each patch) on a large, 75x75 grid
- There are resources of two types: low- and high-value for the player's score
- Players had limited visibility (~5 spaces in any direction)
- The core code-base for the game was derived from Berkeley's Dallinger implementation of this game, modified both to include bots and to be executed "client-side" instead of on a server
- We ran three experimental conditions, as summarized below

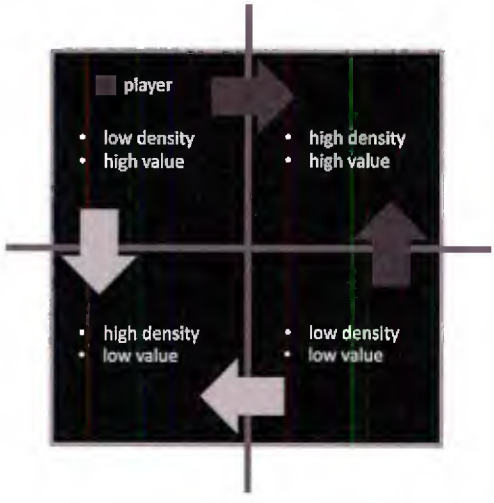
2. Resources in the sensing game



Resource distribution

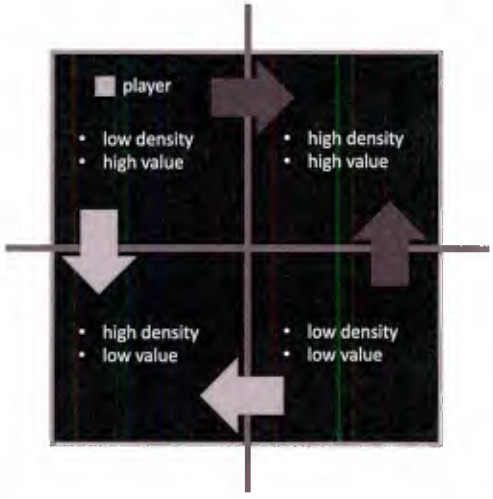
- Patchy resources, with variation in density and value
- All players start in a low density quadrant
- Three experimental conditions based on team social information

3. Experimental conditions



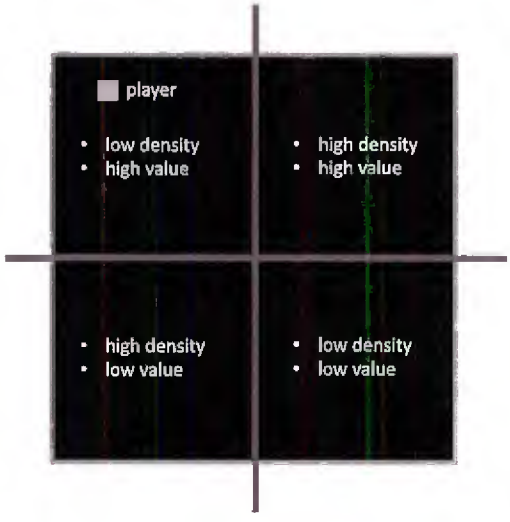
Condition 1: High quality team information

- Player's assigned team are aware of high density, high value patch
- Opponents team are aware of high density, low value patch
- High quality social information from team
- High intra-team competition for best resources



Condition 2: Low quality team information

- Player's assigned team are aware of high density, low value patch
- Opponents team are aware of high density, high value patch
- Low quality social information from team
- High inter-team competition for best resources



Condition 3: No team information

- Neither team is aware of location of patches
- Zero social information from either team
- Low inter- and intra-team competition for best resources

Note the following about these three conditions:

Condition 1: This bot programming induces flocking and strong team-oriented behavior.
Condition 2: This bot programming induces flocking and weak team-oriented behavior.
Condition 3: This bot programming induces no flocking.

4 Pre-registered hypotheses and summary of results

Penn pre-registered four bot-related hypotheses:

a) Human knowledge of bots will decrease rate of human innovation
b) Conformity of bot behavior due to flocking will decrease human innovate
c) Team oriented bot behavior will increase human innovation

d) Bot evolution will increase human innovation.

According to pre-registration, we operationalize a human as innovating if it has a large second-derivative of resource accumulation over time – that is, it is accelerating, over time, in its accrual of resources. (By comparison, bots have only linear resource accrual.) In practice, our analyses involve fitting quadratic functions to the human resource accumulation timeseries, and the value of the coefficient in the quadratic (acceleration) term is the measure of innovation.

We were unable to test Hypothesis D, because technical limitations on client-side Dallinger made it infeasible to implement the sophisticated learning-bot programming before the start of data collection. In practice we were also unable to test hypotheses A, because, in pilot runs, post-game surveys indicated that humans were aware that other players were bots (even if told otherwise) – and so we could not independently manipulate whether humans were aware of bots.

We were able to test Hypotheses B and C – in fact two versions of hypothesis C. The main results are negative: we did not find any differences in rates of human innovation, as operationalized in our pre-registration, between any pair of the three treatment conditions. In particular, we did not find that flocking decreases human innovation, whether or not the flocking is team-oriented; and we do not find that team-oriented bot behavior increases human innovation compared to non-team oriented bot behavior or non-flocking bots.

Although the reproducibility round-robin is limited to PRE-REGISTERED hypotheses only, we have undertaken extensive exploratory research – and we find very significant differences in human behavioral between bot-treatment groups – differences in the spatio-temporal modes of human foraging, in response to the types of bots in the experiment. We believe these are important and actionable results, requiring follow-up experimentation. But they are not pertinent to the reproducibility round robin.

5. Bot programming

There were three types of bot programming: bots unaware of any resource locations, bots aware of the high-density quadrant of high-value resources (purple), and bots aware fo the high-density quadrants of low-value resources (yellow). The bots unaware of any resource location information moved by picking a random space to move towards, and collecting any resources directly along the way, before picking another random location to move to. The bots aware of the high-density resource location quadrant (either the resource of high- or low-value) would pick a random location with the quadrant known to have high-density resources, and move towards that location, with some attraction to other nearby resources encountered along the way. All bot movements included some noise, as well.

6. Data collection

- We collected data from 1,108 unique players over the course 10 days.
- Each game lasted 4 minutes.
- Each game was preceded by a comprehension test and consent form, and followed by a standard questionnaire about gameplay
- Data was collected via MTurk using a modified version of the Dallinger codebase.
- Players were randomly assigned to an experimental condition.
- 126 players were excluded for being inactive (in many cases inactivity results from Dallinger bugs) leaving 981 unique included experiments

7. Raw Data

The raw data are stored in the file "TimeSeriesC2ALL2.csv". These consist of a list of all moves made by all human players across the 1,108 experiments. The CSV file format consists of a single line, with delineations between experiments indicated by "-1".

The human subject in each experiment was assigned to either the Purple or the Yellow team. The raw data on the color of the team assigned to each human in each experiment is saved in the raw data file "ForCl.mat".

The color of the human's team is important – because, for bots aware of resource locations, the Purple bots know the high-value locations, whereas the Yellow bots know only the low-value locations. And so the color assigned to a human's team influences the quality of the information (if any) associated with the bots on that human's team.

8. Processed Data

The raw data for all human moves in all games is parsed by the Matlab script "DalBotRd.m", which produces the parsed output file "DalPlayMov.mat". The parsed output file is a matrix, with each row corresponding to a different experiment, and labelled columns as follows;

St – The strategy of the bots in this treatment. St=1 means the bots in this experiment had knowledge of the high-density resource quadrants (either of low- or high-quality), and thus these bots were flocking. St=0 means the bots in this experiment had not knowledge of resource locations and moved randomly.

Sc, **PosX**, **PosY**, **Tim** – These fields comprise a time-series of the x- and y-position within the grid of the human player, and the cumulative score (Sc), at time Tim

Len - the total number of moves made by the human

ScF – The final score of the human

DTim – The duration of time between the first and last move of the human player.

9. Additional data on bot movements

The raw and processed data described above contain information only about the human subject in each experiment – which is all that is required to test the pre-registered hypotheses. However, the Dallinger session also saves the random seed so that we can reconstruct all bot movements during each game as well, for exploratory analyses.

10. Instructions for reproducing our analyses of the pre-registered hypotheses B and C:

We start by describing all the files in this OSF folder:
"Penn Cycle 2 Report.docx" – this report
"pre-reg.pdf" – the joint UC Berk/Penn pre-registration for cycle 2
"TimeSeriesC2ALL2.csv" – raw data on all human moves in all experiments, described above
"ForCl.mat" – raw data on team assignment of each human, as described above
"DalBotRd.m" – Matlab script to process raw data
"DalPlayMov.mat" – Processed data produced as output of DalBotRd.m
"DallAnal.m" – Analysis script tests pre-registered hypotheses on processed data.

Computational Reproduction Steps:
Step 1: Install Matlab with license, point working directory to include the files above
Step 2: Run DalBotRd.m script to process the data
Step 3: Run DalAnal.m script to test hypotheses

Note that the analysis script fits a quadratic polynomial to the time-series of the human resource accumulation in each experiment, using the poly2 function. The coefficient of the quadratic term is denoted **Acc** in the script. **Acc** is thus the measure of human innovation in each experiment, based on the pre-registered operationalization of innovation.

The following three lines towards the end of the analysis script specifically test our pre-registered hypotheses B and C: In all cases we exclude 126 humans who did not move, as discussed above.

TEST 1, Line 397 (Hypothesis B, for strong team-oriented flocking):
[hacc,pacc]=ttest2(Acc(find(DTim>220 & Len>220 & St==1 & Cl=='P')),Acc(find(DTim>220 & Len>220 & St==0)))
This t-test compares rate of human innovation in a condition with strong team-oriented bot flocking vs innovation with no bot flocking. he result is not statistically significant, p>0.1.

TEST 2, Line 403 (Hypothesis B, for weak team-oriented flocking):
[hacc,pacc]=ttest2(Acc(find(DTim>220 & Len>220 & St==1 & Cl=='Y')),Acc(find(DTim>220 & Len>220 & St==0)))
This t-test compares rate of human innovation in a condition with weak team-oriented bot flocking vs innovation with no bot flocking. he result is not statistically significant, p>0.1.

TEST 3, Line 412 (Hypothesis C, weak vs strong team orientated bots):

[hacc,pacc]=ttest2(Acc(find(DTim>220 & Len>220 & St==1 & Cl=='P')),Acc(find(DTim>220 & Len>220 & St==1 & Cl=='Y')))

This t-test compares rate of human innovation in a condition with weak team-oriented bots flocking vs innovation with strong team-oriented bots. The result is not statistically significant, p>0.1.