Essays on Collective Intelligence

by

Yiftach Nagar

SM Management Research, Massachusetts Institute of Technology, 2013
B.Sc. Industrial Engineering, Tel-Aviv University, 1997

Submitted to the MIT Sloan School of Management
In partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2016

Signature of Author: Signature redacted

MIT Sloan School of Management
April 27, 2016

Certified by: Signature redacted

Thomas W. Malone
Patrick J. McGovern Professor of Management,
Director, MIT Center for Collective Intelligence
Thesis Supervisor

Accepted by: Signature redacted

Catherine Tucker
Sloan Distinguished Professor of Management
Professor of Marketing
Chair, MIT Sloan PhD Program

This page intentionally left blank

Essays on Collective Intelligence

by

Yiftach Nagar

Submitted to the MIT Sloan School of Management on April 27, 2016,
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Management

Abstract

This dissertation consists of three essays that advance our understanding of collective-intelligence: how it works, how it can be used, and how it can be augmented. I combine theoretical and empirical work, spanning qualitative inquiry, lab experiments, and design, exploring how novel ways of organizing, enabled by advancements in information technology, can help us work better, innovate, and solve complex problems.

The first essay offers a collective sensemaking model to explain structurational processes in online communities. I draw upon Weick's model of sensemaking as committed-interpretation, which I ground in a qualitative inquiry into Wikipedia's policy discussion pages, in attempt to explain how structuration emerges as interpretations are negotiated, and then committed through conversation. I argue that the wiki environment provides conditions that help commitments form, strengthen and diffuse, and that this, in turn, helps explain trends of stabilization observed in previous research.

In the second essay, we characterize a class of semi-structured prediction problems, where patterns are difficult to discern, data are difficult to quantify, and changes occur unexpectedly. Making correct predictions under these conditions can be extremely difficult, and is often associated with high stakes. We argue that in these settings, combining predictions from humans and models can outperform predictions made by groups of people, or computers. In laboratory experiments, we combined human and machine predictions, and find the combined predictions more accurate and more robust than predictions made by groups of only people or only machines.

The third essay addresses a critical bottleneck in open-innovation systems: reviewing and selecting the best submissions, in settings where submissions are complex intellectual artifacts whose evaluation require expertise. To aid expert reviewers, we offer a computational approach we developed and tested using data from the Climate CoLab – a large citizen science platform. Our models approximate expert decisions about the submissions with high accuracy, and their use can save review labor, and accelerate the review process.

Thesis Committee:

Thomas W. Malone (Thesis Supervisor)

Patrick J. McGovern Professor of Management,
MIT Sloan School of Management
Director, MIT Center for Collective Intelligence

Jeffrey V. Nickerson

Director of the Center for Decision Technologies
Professor at the School of Business
Stevens Institute of Technology

Iyad Rahwan

Associate Professor of Media Arts & Sciences
Head, Scalable Cooperation Group
MIT Media Lab

Acknowledgments

It really does take a village. I am grateful to so many – teachers and mentors, colleagues, friends, and family – who in countless ways provide me the energy, the inspiration, the drive, the ideas, and the support that allow me to do what I do, and that have helped me complete this part of the journey. First and foremost, my advisor and mentor, Tom Malone, who has been a constant source of support and inspiration – not only in the academic sense, but for life in general – ever since we met on a snowy day that I still remember crisply, several years ago. I'm privileged and proud to be his student.

I'm very grateful to Jeff Nickerson and Iyad Rahwan for serving on my dissertation committee and dedicating the time to provide their support and insightful feedback. Over the years I have had several great teachers and mentors, both at MIT and at Harvard, who have taught me a lot – from statistics to anthropologies of cybercultures; from the tenets of human and machine intelligence to group psychology, and so much more, implicitly. Though they come from very different disciplines, all of them share their contagious passion for teaching, and all have been generous with their attention. I have been very fortunate to learn from Wanda Orlikowski, Erik Brynjolfsson, John Carroll, Paul Osterman, Susan Silbey, Lucy Suchman, Patrick Winston, John Willett, and the late J. Richard Hackman whose sharp wit I dearly miss.

I have also been fortunate to work with and alongside smart colleagues at Center for Collective Intelligence, at the Climate CoLab, and in the greater communities at Sloan, MIT and the area. Rob Laubacher, Laur Fisher, Gary Olson, Peter Gloor, Mark Klein, Bob Halperin, Cris Garcia, Patrick de Boer, Josh Introne, Erik Duhaime, Anita Woolley, Seyda Ertekin, and my fellow students Matt Beane, Avinash Gannamaneni, Shan Huang, Arvind Karunakaran, Heekyung Kim, Xitong Li, Daniel Rock, Guillaume Benjamin Saint-Jacques, Lynn Wu, and Michael Zhao always offered help and advice when I needed them, and lively conversations that enriched me. Alph Bingham, a true gentleman, always found time to chat whenever he came to visit, and every conversation with him has been insightful. I have also learned a great deal from Rob Miller who is always working on cool projects with his students, and I thank him for welcoming me to Crowdfoo.

I thank the organizers and participants of the doctoral consortia at CSCW '13 and SoCS '13, as well as Ezra Zuckerman and Dale Deletis for helpful

conversations at various important stages. There are many others, who have contributed in various ways to the research that is included in this theses, and I included acknowledgements at the end of each essay in recognition of these contributions.

It would have been much harder to navigate MIT without the assistance of Sharon Cayley, Hillary Ross and Sarah Massey from the PhD Program office, and Maria Brennan from the ISO; they always provided prompt support, cheering, and the occasional cozy conversation. The same is true about Richard Hill, whose omni-control of everything one needs to know, stoic smile, and way of making everything seem effortless makes things so much easier.

I'm thankful to the MIT Sloan School of Management, and to the Duwayne J. Peterson, Jr. Fund for funding my PhD fellowship; to the Elie Shaio Memorial fund for award money, and to the MIT Center for Collective Intelligence and its sponsors, who provided financial support as well. The National Science Foundation funded several workshops I attended, as well as several research grants which helped support projects I took part in during my time at MIT. This governmental support and acknowledgment of the importance of unbiased, open science is crucial.

We've been blessed to find friends and to be part of welcoming communities, and that has made a difference. I'm grateful to the teachers at Shaloh house in Brighton, MA, and to Sarah Rodkin and Rabbi Dan Rodkin who have opened their house to us. I'm also grateful to the wonderful teachers at the Devotion school where my children enjoy superb education, and to Elaine Morse who has provided me accommodation and wonderful companionship during my first semester in the program.

Like many PhD journeys, this journey was not short. My friends Danny, Michal, Gili, Ivo, Roi, Shir, Amalia, Bob, Efrat, Dan, Shem, Dorit, Saggi, Hod, Tamar, Eti, and my family, all kept me smiling and are there when I need them, which I cherish. I will be forever indebted to my wife, Ayala, who supported me and encouraged me along the entire way. There is no way I can ever thank her enough.

Like any PhD journey, my journey has started long before I entered the PhD program. I'm forever indebted to my parents, Shlomit and Shaul, who have been my first and most dedicated teachers; who cultivated my curiosity from the day I was born; and who have always been there for me. I can only hope to be able to provide the same for my own wonderful children.

About the Author

Yiftach Nagar is a member of the MIT Center for Collective Intelligence and the MIT Climate CoLab. He explores collective forms of intelligence in sociotechnical settings – from face-to-face groups, to distributed collaborations of minds and machines. His research combines social and cognitive psychology, organization studies and computer-supported-cooperative-work (CSCW) and has been featured in top academic venues, and in the media.

Prior to his doctoral studies, Yiftach's career spanned R&D, project management, and product management roles in several hi-tech organizations. He has led projects and products which were successfully deployed by some of the world's leading telecom companies in Europe, Asia and the US. He holds a B.Sc in Industrial Engineering from Tel-Aviv University and an SM in Management Research from MIT.

Contents

"Over the last decade a new technology has begun to take hold in American business, one so new that its significance is still difficult to evaluate. While many aspects of this technology are uncertain, it seems clear that it will move into the managerial scene rapidly, with definite and far-reaching impact on managerial organization. In this article we would like to speculate about these effects, especially as they apply to medium-size and large business firms of the future. The new technology does not yet have a single established name. We shall call it information technology."

— Harold J. Leavitt and Thomas L. Whisler, Management in the 1980's,
Harvard Business Review 36:41 (Nov.-Dec.) 1958

"...there is nobody here but us scratching around trying to make our experience and our world as comprehensible to ourselves in the best way we can, that the various kinds of order we come up with are a product of our imagination and need, not something dictated to us by Reality Itself. There isn't any One True Map of the earth, of human existence, of the universe, or of Ultimate Reality, a Map supposedly embedded inside. these things; there are only maps we construct to make sense of the welter of our experience, and only us to judge whether these maps are worthwhile for us or not."

— Brian Fay (1990). Critical Realism?
Journal for the Theory of Social Behaviour, 20(1), 33-41

Thesis Introduction

Our species' ability to cooperate and collaborate at scale has been a major determinant – arguably the most important one – of our prosperity. Throughout history, advancements in information and communication technology, from cave drawings, to the development of language, writing, printing, and electronic media, have allowed people to communicate, coordinate, and collaborate at growing speed, scale, and complexity, accelerating the rate in which ideas are disseminated and exchanged, fueling innovation, and strengthening our competitive advantage.

The rate and impact of these changes have been especially significant during the so-called 'Digital Revolution' of the second half of the 20th century. But over the past two decades, we are seeing even more radical shifts, largely affected by the ripening of two sets of technologies: the Web, and Artificial Intelligence. The Web (together with the host of technologies that underlie it, and rely on it) has enabled people to connect in unprecedented scale and speed. It facilitated the creation of new types of organizations such as Wikipedia, citizen-science projects such as FoldIt (where novices help scientists to discover protein structures), and many other, which brought together hundreds of thousands of volunteers to work on important, large-scale projects. More recently, advancements in Artificial Intelligence (and related technologies, including distributed and parallel computing, robotics, etc.) have led to major leaps in the ability of computers to reason and act in the real world. In the past few years, AI has demonstrably equaled and surpassed human performance in a growing number of complicated tasks. These advancements are already affecting not only the job-markets of low-skilled workers, but also higher up the chain.

Yet, most of our organizations still structure themselves and operate in traditional ways; we are only beginning to understand the potential of these new opportunities. Indeed, while the introduction of new technologies often provides us an occasion to re-examine the ways we do things, and the opportunity to invent new, better ways of doing them, history shows that it can take some time until the potential of new technologies – especially 'general purpose' technologies – is fully realized. For instance, when electric motors became available to American factories during the late 19th century, they were initially used directly in place of the steam engines that powered factories, which led only to limited savings. It took about three more decades to realize, develop, and implement new factory layouts and work processes that took full advantage of the modularity enabled by the new technology, which ultimately led to huge gains[1].

How then, can we organize better, to take advantage of the new opportunities that new information technologies (specifically: the web, and artificial intelligence) provide us? In what ways should we connect together experts, crowds, and computers, in order to cultivate their collective intelligence and address big challenges?

---

[1] Devine, W. D. (1983). From Shafts to Wires: Historical Perspective on Electrification. The Journal of Economic History, 43(2), 347-372.

Advancing a Research Agenda for Studying Collective Intelligence

In the ecosystem of inventing, evolving, and stumbling upon new ways of organizing, purposeful, systematic research can help us more quickly find, understand, develop, and converge on new and better modes. I have focused my work on two lines of research that aim to advance aspects of this agenda: on the basic research side, understanding collective intelligence & sensemaking: how do they work? what supports them? what hinders them? which I do through analyses of group processes; on the applicative research side, I explore ways of *using* collective-intelligence (e.g. by connecting people and computers to improve task performance); and ways to *enhance* collective-intelligence, through design and experimentation.

Understanding Collective-Intelligence

In Essay 1, I follow the collective-sensemaking process through which the Wikipedia community negotiates and constructs its policies. I analyzed conversations that took place over the course of months as members of the community negotiated conflicting views, assumptions and desires, and managed to develop *committed interpretations.* Drawing on the literature of collective sensemaking in organizations, I offer a process model which I ground in the analysis, and which explains how structuration emerges as interpretations are negotiated and then committed through conversation, and as they are reified in the policy. I further propose that the wiki environment provides conditions that help commitments form, strengthen and diffuse, and that this, in turn, helps explain trends of stabilization observed in previous research. The model may prove useful for understanding collective-sensemaking processes in

other large wiki communities, and potentially in other radically open organizations.

During this study, I noted that a great part of the conversation comprises of questions (and answers). This led me to form research questions and hypotheses regarding the role of questions as a device through which we engage others in collective thinking, which I am interested to pursue further in the future.

Using Collective-Intelligence

In order to gain from the advancement of information technologies and enhance the collective-intelligence of groups, organizations, and society, it is helpful to consider their new affordances, ask what we can do now that we could not do as well (or at all) before, and engage in design, prototyping and experimentation. Specifically, I am interested in ways of gaining from combining human and machine intelligence, and in ways of connecting diverse groups of people – including experts and novices – in order to address big challenges.

Combining Human and Machine Intelligence for Making Predictions

The ability to make predictions is one of the foundations of intelligence, and intelligent systems. For organizations, and managers, predicting future events is a crucial task. How can we make better predictions? What can we gain from connecting people and computers?

Prior research has shown that in relatively structured, stable environments, statistical models are almost always at least as good as human experts at making predictions, and often substantially better. However, many important prediction problems in the real world arise in

semi-structured environments where data are difficult to codify or quantify, where patterns are difficult to discern, and where changes occur unexpectedly. In Essay 2, we hypothesize that in these environments, where it may be difficult – or even impossible – to build reliable and dependable predictive models, combining people and computers can lead to improved predictions.

To test this hypothesis, we conducted laboratory experiments in which we used prediction markets, human judgment, and averaging to combine predictions from groups of people and artificial intelligence agents. We find the combined predictions both more accurate and more robust than predictions made by groups of only people or only machines.

Enhancing Collective-Intelligence

Connecting Experts, Crowds and Algorithms in Open Innovation Systems

In a wide variety of domains, organizations increasingly turn to crowds in search for novel ideas and solutions. Indeed, the advantages of involving a large and diverse crowd in creative problem solving are both theoretically sound and empirically supported. However, for tackling complex challenges, it is often not enough to "throw the problem out there". Working with the crowd often requires preparation and translation work, and central guidance from experts. This has also been our experience at the Climate CoLab (www.climatecolab.org), a citizen-science project initiated at the MIT Center for Collective-Intelligence, which connects experts with a diverse community of people from all over the world, in order to develop plans for addressing climate change. Specifically, I have been focusing on issues related to the review of complex intellectual artifacts.

Improving the review of complex intellectual artifacts: addressing scale

In the context of open innovation and crowd ideation, with the blessing of receiving many more ideas, comes the challenge of evaluating those ideas. In the CoLab, we have been receiving hundreds of submissions every year. Some other challenges draw thousands of ideas, and more. Extreme cases, such as BP's crowdsourcing effort following the DeepWater Horizon disaster, or Google's 10^{100} contest, drew over 120,000 entries each. With a limited number of experts able and available to review entries, the review bottleneck poses a real barrier to innovation, and can cause significant delays.

I interviewed judges and studied the literature of expertise, and of peer-review. I talked with many other people involved in the evaluation of complex intellectual artifacts – conference paper chairs, NSF officers, operators of open innovation platforms – to understand the magnitude of the problem, and to learn of the ways different organizations cope with this barrier. I learned that the problem was real, common, and important

In Essay 3 we describe a mechanical approach to aid expert judges by computationally rating submissions. Developed and tested with data from innovation contests that were held in the Climate CoLab, our model achieved encouraging results: if we use the model in the most conservative way, maximizing sensitivity, we can skip the review of ~15% of the submissions. Allowing ~10% false-negatives (which can be handled by a secondary review process by non-experts), we can potentially save the experts the need to review ~50% of the submissions. Our approach and many of our metrics can be adapted to other settings, and I intend, with my collaborators, to keep developing and testing it using datasets from additional settings.

Essay 1

Understanding Collective-Intelligence:

The Structuring of an Online Community as a Collective-Sensemaking Process

Yiftach Nagar[2]

Abstract

I observe conversations that take place as Wikipedia members negotiate, construct, and interpret its policies. Logs of these conversations offer a rare – perhaps unparalleled – opportunity to track how individuals, as they try to make sense, engage others in social interacts that become a collective processes of sensemaking. I draw upon Weick's model of sensemaking as committed-interpretation, which I ground in a qualitative inquiry into policy discussion pages, in attempt to explain how structuration emerges as interpretations are negotiated, and then committed through conversation, and as they are reified in the policy. I argue that the wiki environment provides conditions that help commitments form, strengthen and diffuse, and that this, in turn, helps explain trends of stabilization observed in previous research. The proposed model may prove useful for understanding structurational processes in other large wiki communities, and potentially in other radically open organizations.

---

[2] This is a slightly edited version of a paper that was originally published as:
Nagar, Y. (2012). *What Do You Think: The Structuring of an Online Community as a Collective-Sensemaking Process.* In proceedings of the 2012 ACM Conference on Computer Supported Collaborative Work (CSCW '12), Seattle, WA, USA. 978-1-4503-1086-4 http://dx.doi.org/10.1145/2145204.2145266

"A dominant question for scholars of organizing is: How do people produce and acquire a sense of order that allows them to coordinate their actions in ways that have mutual relevance?"

(Weick, 1993)

"Wikipedia did not arise spontaneously; it arose through people interacting and, as a result of that interaction, finding ways that worked."

Interviewee 3 (I3) (Forte & Bruckman, 2008, p. 2)

Introduction

How groups, organizations, communities and societies form and change over time has been a key subject of inquiry in the social sciences. The structurational perspective (Barley, 1986; Orlikowski, 1996) posits organizational transformation as endemic to the practice of organizing, embedded in, and emergent from the situated daily practices of organizational members – "an ongoing improvisation enacted by organizational actors trying to make sense of and act coherently in the world" (Orlikowski, 1996, p. 65). In this paper, my goal is to track such efforts of sensemaking by people in one social system, to learn about how these efforts lead to structuring of the organization.

The organization chosen as the site for this inquiry is Wikipedia, for a number of reasons. The first is opportunistic: studying real social settings, 'in the wild' often entails spending relatively long periods of time in the field, especially for researchers who are interested in collecting micro-level data. Also, ethnographers and other social scientists who pursue micro-

level data, are bound to miss a lot of what is going on, as they are limited, physically, to being at certain times and places. The discussion archives of Wikipedia provide us unique access to vast amounts of verbatim conversations among its members. Because most interactions among Wikipedians are done online and remain documented, we get to glimpse into Wikipedia's communal and organizational stream of consciousness.

Second, the shape of social interaction inside many organizations is changing as interpersonal communications are gradually shifting weight to textual interactions over social-software platforms. Wikipedia is one extreme example, as interaction among its members is almost entirely public. It therefore not only provides opportunity to study micro-level interactions among members of *an* organization, but rather, to do so in web-based organization that is radically open and distributed. Assuming more organizations in the future will share at least some of these qualities (Malone, 2004; Malone, Laubacher, & Scott Morton, 2003) (even if to lesser extent in comparison to Wikipedia), it is interesting to learn about how they might work, and to see whether, and to what extent, theories developed to explain more traditional organizations hold, and what assumptions might need revision. Finally, Wikipedia in itself has drawn focus of a diverse community of researchers, and this study can enhance our knowledge of some aspects of its work – specifically, to highlight some processes through which parts of its structure are formed and transformed.

I therefore set out to explore the ongoing process of sensemaking Wikipedians conduct as they discuss, negotiate, construct and change one of its policies. As detailed later, I took a grounded approach to this qualitative inquiry, while leaning on ideas from Karl Weick's work on

sensemaking in organizations (Weick, 1993, 1995) in interpreting and explaining the findings.

Background

The Role and Importance of Wikipedia's policies

Wikipedia's success (as measured by several parameters, including popularity, engagement, and quality of its articles) has surprised many skeptics, and has been widely discussed. Wikipedia has organically developed a complex bureaucracy, which includes an organizational structure, organizational processes, and many formal "objects", including policies, guidelines etc. Several researchers point to the important part policies, rules, and guidelines play in Wikipedia's daily operation and their contribution to its success (e.g. Beschastnikh, Kriplean, & McDonald, 2008; Bryant, Forte, & Bruckman, 2005; Butler, Joyce, & Pike, 2008; Forte & Bruckman, 2005; Kriplean, Beschastnikh, McDonald, & Golder, 2007; Morgan & Zachry, 2010; Viégas, Wattenberg, & McKeon, 2007).

Policies deal with a wide range of contexts – from matters of content, to rules of proper conduct, to discussion of enforcement and more (cf. Beschastnikh et al., 2008). Thus, they help Wikipedians make sense of complex situations and serve as references to legitimize action (Beschastnikh et al., 2008; Kriplean et al., 2007).

Policies are not merely prescriptive of social behavior. Wikipedia's policies and guidelines (and all other components of the bureaucracy) are developed by the community in attempt to capture and institutionalize best practices. What is considered best practice is a matter of consensual

view, and it is expected that any proposed change should usually be discussed in advance *"to ensure that the change reflects consensus"*[3]. Thus, policies and guidelines are also reflective of social practice (see also Forte & Bruckman, 2008).

The accounts discussed above have helped us gain insight into the role of policies in regulating ongoing activity in Wikipedia. What has been less discussed, however, is the *process* by which the bureaucracy emerged, and specifically, how the policies are formed and transformed. Forte and Bruckman (2008) dedicate parts of their discussion of Wikipedia's distributed governance structure to the process of policy creation, but the theoretical lens guiding their inquiry is a sociological one, drawing on theories of commons-based governance, and accordingly, their focus is the organizational environment and setting rather than the cognitive processes and micro-level interactions that lead to the social construction of the policy. Elaborating a case study of the creation of one of Wikipedia's policies, they carefully examine at the *"thick tangle of circumstance"* that set the stage for the process of creation of what started as a guideline, and later became a policy, and summarize: *"Eventually, after much off- and on-wiki discussion about the situation, a proposal page was started and the community began constructing what was initially a proposed guideline. Eventually, the page reached a form acceptable to most community members"*. What they leave unexplained, when they write *"Eventually"* (twice!), is exactly this *"much of- and on-wiki discussion about the situation"* and the social process of construction of the

---

[3] http://en.wikipedia.org/wiki/Wikipedia_policies, accessed 2010-5-15

proposed guideline until it *"reached a form acceptable to most community members"*.

In this paper I take a close look at the discussions among Wikipedians as they struggle to make sense of their social reality and reach consensus. I posit these discussions as a collective process of sensemaking, and, drawing on Weick's concept of ***committed interpretation*** (Weick, 1993), propose a model of how social structures within Wikipedia might emerge from this process. Before discussing the research, I briefly introduce some ideas about sensemaking in organizations.

Sensemaking and Committed Interpretation in Organizations

The study of sensemaking in organizations has produced tremendous amounts of work, which cannot be reviewed here. For the purposes of this paper, it is useful to highlight just a few points regarding sensemaking that will provide a substrate for discussion.

What is the study of sensemaking in organizations? The following points are drawn, adapted, and synthesized mainly from Weick (1995), and, due to limits of scope, are brought here only in summary form[4]. Three – interrelated – key questions for researchers of sensemaking in organizations are the following: *"How are microstabilities produced in the midst of continuing change? How do people produce and acquire a sense of order*

[4] Interested readers would find a much more subtle, elaborate and nuanced discussion in Weick (1995). The discussion in chapters 1, 2, 6 and 7 is especially insightful and related to the ideas discussed here.

that allows them to coordinate their actions in ways that have mutual relevance?" (Weick, 1993), and, *"how are meanings and artifacts produced and reproduced in complex nets of collective action?"* (Czarniawska-Joerges, 1992; cited in Weick, 1995, p. 172).

Sensemaking has to do with interpretation, of course, but interpretation is just one component of sensemaking, which is also concerned with the construction of reality. Sensemaking is enactive of sensible environments. What this means is that as people try to make sense of reality, they do more than trying to cope with entities that already exist in the world and interpret them: in organizational life, as people act, they create materials and settings which then become constraints and opportunities in the environment they face.

Sensemaking can take many forms and work in many ways. In this paper I use one prototype of sensemaking in organizations – *Committed Interpretation* – that is offered by Weick (1993) as a possible answer to the first question mentioned above (and, which, I believe, helps deal with the other two questions as well). In concise form, *"The concept of committed interpretation suggests that people become bound to interactsrather than acts, that the form of interacts is itself committing, and that justifications of commitment tend to invoke social rather than solitary entities. These three seeds of social order enlarge and diffuse among people through enactment, imitation, proselytizing, and reification, thereby imposing order on confusion"* (Weick, 1993).

In the following sections I elaborate this concept, ground it in data collected from policy discussion pages, and show that it can help us understand and explain the processes of structuring of Wikipedia, beyond what was offered in previous accounts.

Tracking the Ongoing Processes of Policy Structuring

Method

I focused my inquiry on Wikipedia's "Neutral Point of View" policy (NPOV), which is one of its core content policies (arguably, the most fundamental), and one which has drawn a lot of attention, discussion and action, as I detail below. In a nutshell, this policy states: "Articles mustn't *take* sides, but should *explain* the sides, fairly and without bias. This applies to both what you say and how you say it". As Reagle (2005) points out, *"...in the Wikipedia culture, the notion of 'neutrality' is not understood so much as an end result, but as a process"*[5], *and this policy details parts of this process.*

The centrality and importance of NPOV to Wikipedia and its special constitutional status are captured in the following remark by Wikipedia's co-founder and de-facto leader, Jimmy Wales, excerpted from a talk he gave in 2005:

---

[5] While I agree with Reagle that this is the original intent of Wikipedia's founders and its core elite, based on my observations I believe that not all editors understand this difference, or accept it. This discrepancy is one source of fuel for the continued discussion of NPOV, at least during the period I have observed.

> *"So how do we do this? [...] how does it work? So there [are] a few elements, mostly social policies and some elements of the software. So the biggest and the most important thing is our neutral point of view policy. This is something that I set down from the very beginning, as a core principle of the community that's completely not debatable." (Wales, 2005)*

It is no wonder, therefore, that this policy is highly visible and serves as reference in many discussions. Ironically, what was proposed as a simple, "completely not debatable" core principle is highly discussed even today, a decade after the founding of Wikipedia. While the rate and intensity of the discussion varies, it is yet to reach an asymptote or level, and it may possibly continue indefinitely. According to *WikiChecker* (http://en.wikichecker.com), between April 2006 and May 2010, more than 800 people have participated in the online discussion of the NPOV policy page itself (i.e. in the associated "talk" page), and performed about 9500 edits of that page. Note that these discussions are assumed to revolve mainly around issues of framing the policy itself, whereas discussions of interpretation related to enforcement of the written policy are supposed to take place elsewhere (mainly over the Neutral point of view Noticeboard and in administrator discussions). The policy itself has been edited over 4,500 times between February 2002 and November 2011 (more than once a day, on average), by more than 1700 people.

I focused my attention on a period ranging from July 2005 to January 29, 2006 (NPOV discussion archives 004 – 014), as that period has produced profuse discussion (11 archives for a period of 7 months, out of 37 archives for the period 2004 – February 2010). I coded about one quarter of those 11 archives, and also sporadically sampled some NPOV discussion pages from earlier and later periods. I started with line-by-line

coding (Charmaz, 2006, pp. 50-53) which was relatively open at first, and then gradually moved to coding larger fragments in a more focused manner. With the goal of tracking individual and collective acts of interpretation, sensemaking, intelligizing and construction, informed mainly by discussions of sensemaking (Weick) and structuration (Barley, Orlikowski, and others), I looked for expressions of surprise, puzzle, questions, clarifications, agreement and disagreement, divergence and reconciliation, and other expressions that related to sensemaking and interpretation. I augmented this by coding additional pages of the discussion around Wikipedia's "Five pillars" page[6], and have reviewed several other auxiliary materials including interviews with Jimmy Wales, correspondence of Larry Sanger (co-founder of Wikipedia and its first community discussion leader), and many other online resources, and also used tools e.g. Wikidashboard (Suh, Chi, Kittur, & Pendleton, 2008), and Wikichecker, that helped me navigate Wikipedia and make sense of what I see.

Committed Interpretation in Wikipedia

In this section I present a thick description of the processes of sensemaking I have observed, grounding the discussion of 'committed interpretation' in the data.

---

[6] This is a relatively new page, which is an attempt to create a higher level framework – a constitution of core principles out of which other rules are derived. This page's necessity, its status, and its content have all been subject to debate.

Committing to Interacts

Sensemaking, as most action, is inherently a social phenomenon. As people try to make sense, they interact with others, whether those others are present in the moment, or imagined, because people know their actions and explicit interpretations will have to be understood, accepted, and implemented by others. Therefore, Weick argues, when people become bound to acts, those acts tend to be interacts rather than solitary acts. Further, in social settings, actions that are public, irrevocable, and volitional are harder to undo and disown, and therefore, create commitment. Therefore, Weick concludes, *"interacts themselves generate their own conditions of commitment since each party's action is public, irrevocable, and volitional relative to the other party in the exchange"* (Weick, 1993, p. 20).

In Wikipedia's discussion pages, every act – and therefore, also, every interact - is indeed public, irrevocable (as the history of edits, including that of discussion pages, is kept), and of course, volitional. The Wiki environment, as a medium for interaction, therefore provides conditions that serve as catalysts for turning such interacts to commitments, because each party's action is visible not only to the other party in the specific exchange, but to anyone (both inside and outside the community), for an indefinite time.

Interaction over the discussion pages takes form in various rhetorical acts. The main form I have observed involves posing questions and proposing answers.

Posing Questions

I have identified several archetypes of questions that people ask, including informative questions, provocative question, etc. Notably, editors ask a lot of questions in attempt to make sense of others' views. I coded those as follows:

1. Asking clarification questions

For example:

> *Causa, why do you say there is no need for an introductory sentence? If we have no sentence and no tag (and it's unlikely that any of the standard tags fit this page, so tagging in this case effectively means putting the introductory sentence into a box), then people coming to this page won't know what its purpose or status is. Why would we want to add to the mystery here? (That's not to say someone might not improve the introductory sentence we currently have.)*
> *<<Kotniski (talk) 08:56, 7 November 2009 (UTC)*

As seen in that quotation, clarification questions are often asked not only to understand an issue about the policy itself, but rather, in order to understand *what someone else thinks* about that issue. In these cases, the question may sometimes be explicitly addressed to a specific person (e.g. "Causa, " above), but many other times it is clear from the discussion that a dialog or a conversation develops organically between two or more people as in this example:

I agree with John that a 're-shuffle' of relative importance of policies and guidelines, as originally proposed by TMoW, is probably not the best way forward (note that I propose some 'precedence-reshuffle' every once and a while myself, but that's a very slow moving process, best you know that)
--Francis Schonken 21:17, 9 October 2005 (UTC)

Thing is, some POVs aren't worth including at all. But how to distinguish? As for NOR7, i mean prove your point on the talk page (with ext sources) to see if it worth including. This is what i mean by having a reliable claim before its elevated from "justanother-claim" to another "POV". As John says, there are lots of editors which include any ol' claim to maintain NPOV. And once you decide to include, then by what degree and how? I feel a clarification/specification would clear up a lot. Okay so we kind of agree. But now what?
--The Minister of War 05:54, 10 October 2005 (UTC)

Again, in this excerpt we see that the user named "The Minister of War" is interested in knowing what his peers think. Thus, implicitly, he (she?) does not perceive the policy as a "fixed object" that is "out there". Rather, the policy is viewed as what he and his peers decide that it is - what they *make* of it. Asking "But now what?" is an attempt to elicit a proposition for action, and to continue the conversation, by which the interact will become a double interact, etc., and commitments will grow.

---

[7] Wikipedia's policy on "No original research".

2. Asking about behavior, trying to understand the rules

Such questions are usually asked by people not sure what to do in various cases. Technically, the policy discussion pages are not the 'right' place to do this. However, since the bureaucracy is so overwhelming, some people are just not sure where they should channel their questions. These are not "total newbies" usually, since those are not very likely to reach the policy discussion pages. Thus, although not 'appropriate', discussions about policy use and enforcement sometimes blend into the discussion about the policy construction. Here are two examples:

> *I found some questionable pov elements from an article on John Milius and added a check pov template and removed the questionable elements. There has not been any response on the talk page nor any further revisions. Do i take down the check pov template? how long do i have to wait?*
> *-Seasee 22:10, 21 Jun 2005 (UTC)*

> *Some articles use terms like "God", "white people", "luck", "Jewishness", "fairies", "nobility" that people have invented to support various religious/superstitious or political programs. If I don't believe any of this stuff do I have a POV? Should I insert "so-called" or "alleged" in front of these terms?*
> *24.64.166.191 06:03, 9 Jun 2005 (UTC)*

All these types of questions demonstrate difficulties in interpreting the meaning of the policy. What people are actually doing by asking them, is trying to make sense, and their way to try and make sense is to engage others in conversation.

Engaging others in conversation is not the only way to try and make sense. One could, for example, read more. Obviously, some people may prefer to read more in order to try to make sense, while others prefer to ask. Even if more people prefer to try other methods first, eventually, there are at least some people who pose opinions and questions on the policy discussion page. This engagement of others in the hermeneutic process is how the personal act of trying to make sense of things first becomes an interact (by asking a question, and receiving answer); and then, the interact becomes committing – as the person either accepts the answer (the simple case), or resists, which may lead, through a longer process, to a change. Individual attempts to make sense thus become a collective process of sensemaking. Through this process, participants (both active and "passive", i.e. those who read the conversation of others) gain a better understanding of what others think/feel, and of other facets of their environment (temporary and tentative as those may be), and interpretations are created, negotiated, modified and committed. The policy, then, is the manifestation of these commitments.

3. Asking questions as a rhetorical technique

A different class of questions comes up, which are part-question, part-suggestion. When editors wish to propose a change in the policy, they sometimes do it by posing a question. This is done in the spirit of Wikipedia which strives to achieve consensus, and so, before actually being bold and introducing the change, they ask others for their opinion. For example:

> *There are places in the Wikipedia namespace where advice and guidelines are offered (as distinct from policy), and while a majority of Wikipedians may support this advice, there may be examples where a significant minority disagree (I have in mind inclusionist/deletionist type dichotomies). In such cases, should the NPOV policy be read as to force the inclusion of strong minority positions (e.g. ~20% support) into pages that discuss Wikipedian behavior (e.g. Offer voting guides).*
> *Dragons flight 20:47, Jun 19, 2005 (UTC)*

Asserting by asking is another way of using questions as a rhetorical "trick". The person asks a seemingly technical question about something perfunctory and at the same time makes a statement by taking something as fact. For example, the following question:

> *Where would it be appropriate to add a blurb about quotation marks being used as a form of bias?*

is seemingly technical - but at the same time also states that quotation marks are a form of bias.

Rhetorical questions and provocative questions are also occasionally posed, to promote or challenge an ideology or point of view. Sometimes, this can trigger further discussion, as in this case:

According to Wikipedia founder Jimbo Wales, NPOV is "absolute and non-negotiable". How is this different from religious dogma? It seems to me NPOV is taken to absurd extremes by some self-important administrators too enamored of their little barnstar award trinkets. Not to mention it is overly sanitizing the pedia to where even a sense of humor becomes verboten, hence the need for the new admin award category: the award of the NPOV Nazi
(unsigned, undated)

Well, you do have to give some creedence to this argument. Whenever anything is taken to be an absolute, it opens the door for a person to manipulate that to their own ends. Let me give you an example: the Salem Witch Trials. All someone had to do was describe someone as a witch, and then they could freely assualt that person with no consequences. Well, isn't it possible (indeed, likely) that someone might take this sacred cow, NPOV, and accuse someone else of being not NOPV simply for the ability to attack them or their ideas free of consequences? Surely, there must be safeguards to protect those who are the unortunate victims of this sort of manipulation. May I ask what provisions you have thought of to prvent NPOV from becoming a tool for witchhunters?
Dave (undated)

But note: even though rhetorical questions are not directly looking for information, a rhetorical act is one whose purpose is to persuade others. In that, rhetorical questions, too, are a mechanism by which people attempt to engage other minds, offer and seek interpretations, and pursue common grounds.

Proposing Answers

By answering questions posed on the discussion pages, I find that Wikipedians are doing several things, beyond the sharing of 'dry' information:

1. Offering interpretation.

Often (and especially when discussing such a term like NPOV, which is loaded with ambiguity and possible interpretations), questions are not simple informative questions, but are posing principle challenges of interpretation. Questions and answers serve as mechanism for a social process of hermeneutics. Questions of interpretation come up in the discussions of people who actually enforce the policy (which I do not discuss here), but I also found some traces of these discussions in the policy discussion pages. Indeed, it is often difficult, and perhaps impossible, to agree on a common interpretation. This difficulty is evident in the vibrant discussion of NPOV, as evident both in many of the actual answers I have observed and in the numbers that point turbulent editing of the policy, and which can also be inferred from the following note of Butler et al. (2008): *While the "Ignore all rules" policy itself is only sixteen words long, the page explaining what the policy means contains over 500 words, refers readers to seven other documents, has generated over 8,000 words of discussion, and has been changed over 100 times in less than a year.*

In this context, it appears that the concept of *committed* interpretation should be preferred over that of *shared* interpretations, or *shared* meanings, as it alludes to the *satisficing* character of the activity of people (see also Weick, 1995, pp. 42-43 in that regard). Meanings and interpretations are never shared by all the people, all the time, under all

circumstances. They are always only temporarily shared, never in whole, never by all. But while it is not always possible to have shared interpretations, it is still possible to find common ground even without them, by finding such interpretations to which people can commit. These may at least allow action to proceed.

Similar observations are made by Brennan as she discusses grounding in conversations: *"Understanding is not the same as agreement or uptake. When speakers and addressees have incompatible intentions, they might understand one another perfectly well but 'agree to disagree'" (2005, p. 125); "Grounding Is Only as Precise as it needs to be [...] people in conversation do not try to get their hypotheses to converge perfectly-in fact, since neither party is omniscient, this is not even feasible. Instead, they try to reach a level of convergence that is sufficient for current purposes, satisficing in Simon's (1981) terms. "(ibid., p. 104)*

Thus, it should be clear that committed interpretations are only temporary points of stability in space-time, sensitive to change in circumstances. As circumstances change, commitments can be revisited, and broken.

2. Explaining and signaling to others what they think is the answer to the question.

Because NPOV deals with such fundamental issues that touch epistemological and even ontological issues and eventually boil down to *beliefs*, answers serve not only to express an opinion. By *signaling* I mean that, eventually, whether formally or not, Wikipedians do hold tallies of voices. Consensus is a key value in Wikipedia, but majority voices are counted more than minority voices (in fact this is a part of NPOV itself). If more people support an opinion - even in a policy - this opinion is more

likely to be represented. I found this sort of signaling in another type of contribution to the discussion which is not technically an "answer" and which I coded as "Seconding input from another member" or "supporting proposition". Sometimes people would write things like "I agree" or "I agree with X" or similar. Indeed, at some point in time a norm of stating one's opinion in one word – "Agree" or "Oppose" (or similar words) – followed by further explanation, has emerged in policy discussion pages. By doing so, people clearly indicate their commitment to the proposed interpretations.

3. Explaining to themselves.

The mere act of writing helps people construct their arguments. By choosing to answer, people engage in a committing interact. It is easy to envision a case where a person would answer a question, and then be dragged, against her/his will into a longer dialog, and indeed such cases appear frequently in the discussion pages (with some comments expressing weariness and loss of patience). But once the conversation has started, it is not easy to disengage from it, at least not without 'losing face'. When it comes to persuasion, disengaging may have a price that is higher than that of not engaging in conversation in the first place, as it may be perceived as "admitting" to being wrong, or to accepting a certain opinion with which the editor does not really agree.

Thus, the entire discussion - questions and answers, may lead, eventually, to the forming of new interpretations that lead to changes to the policy, or, in other cases, serve to and re-enact the policy.

Invoking Policies to Justify Commitment

As mentioned above, policies and guidelines help Wikipedians make sense of complex situations and they are widely used by Wikipedians as references to legitimize action (Beschastnikh et al., 2008; Kriplean et al., 2007). For example, Buriol et al. (2006) mention that the "3-revert-rule" policy which was introduced in response to a growing number of "edit-wars" (recurring reverts by two sides arguing) has had an immediate effect of decreasing occurrences of those "double reverts". As one Wikipedia editor noted:

> *"The degree of success that one meets in dealing with conflicts (especially conflicts with experience[d] editors) often depends on the efficiency with which one can quote policy and precedent." (Kittur, Suh, Pendleton, & Chi, 2007)*

These findings are in accord with the third part of the definition of committed interpretation, namely that *justifications of commitment tend to invoke social rather than solitary entities*. My data collection focused on policy discussion pages, where the discussion mainly concerns the *construction* of the policy, rather than its *use*, and therefore references to policies seem to appear somewhat less frequently, but they certainly do appear, as do references to other social entities such as Wikipedia's Arbitration Committee (ArbCom)[8].

---

[8] The ArbCom is "a panel of editors that imposes binding rulings with regard to disputes between other editors". See http://en.wikipedia.org/wiki/Arbitration_Committee_(English_Wikipedia)

Beschastnikh et al. (2008), find significant growth in policy citations over time. They also find that enforcement (as manifested in policy citations) has diffused into the larger body of registered users, with the practice of policy citation increasingly becoming commonplace. Similarly, Butler et al. (Butler et al., 2008) and Forte et al. (Forte, Larco, & Bruckman, 2009), note that over time making changes in the policies has become more difficult, and has slowed. Weick's conceptualization of the process of committed interpretation helps explain for these observed phenomena as well as it argues that what he recognizes as the "*three seeds of social order" (namely:* that people become bound to interacts rather than acts, that the form of interacts is itself committing, and that justifications of commitment tend to invoke social rather than solitary entities*)*, "enlarge and diffuse among people through enactment, imitation, proselytizing, and reification, thereby imposing order on confusion". Invoking policies (and other parts of the bureaucracy) in the discussion as a means to justify commitment serves to reinforce them. As policies are enacted by administrators, and invoked as justifications in discussion, they become reified, and commitments grow. This creates a positive feedback loop, where enactment and reification of the policy feed each other.

As this "evolutionary" process makes changes to the policy rarer, editors redirect their efforts to creating and updating less formal parts of the bureaucracy, such as guidelines and essays (Morgan & Zachry, 2010).

The Seeds of Change

Weick asks how people produce and acquire a sense of order that allows them to coordinate their actions in ways that have mutual relevance. His answer is: by concrete communicative interaction in which

people invoke macro structures to justify commitments. He concludes: *"Thus, social order is created continuously as people make commitments and develop valid, socially acceptable justifications for these commitments. Phrased in this way, individual sensemaking has the potential to be transformed into social structures and to maintain these structures. Commitment is one means by which social structure is realized. This proposal suggests a possible mechanism by which structuration (e.g., Barley, 1986; Giddens, 1984) actually works."*

Indeed, the story I told so far, is mainly a story of regulation and creation of stability. But what about organizational change?

Structuration theory offers a dialectical, reciprocal account of social change, and has been adopted and adapted by organizational researchers to explain organizational change. It posits that social structures enable and constrain the actions of agents, and yet, do not determine their actions. Several notable works have tracked structuration processes following an external 'shock' such as the introduction of new technologies (Barley, 1986; Orlikowski, 1996), or new regulations (Kellogg, Forthcoming) into an organization. Although in all these works change starts with an external shock, Orlikowski highlights the notion of change as an ongoing improvisation and quotes from March that "in its fundamental structure a theory of organizational change should not be remarkably different from a theory of ordinary action" (March, 1981, p. 564; quoted in Orlikowski, 1996, p. 66). She further locates the beginning of the change process in the attempt of people to make sense of *a new situation.*

Obviously, when a big, external shock is introduced, it is not surprising that people try to make sense of it. But, excluding external shocks, what

causes a new situation during 'ordinary action'? and how, exactly, are people attempting to make sense?

As seen in Wikipedia's policy discussion pages, collective-sensemaking can also start with a single person reviewing or reflecting previous understandings without any evident external trigger.

In Wikipedia, every newcomer may introduce an 'occasion' to discuss and negotiate meaning. This might be somewhat different in traditional organizations. There, if a newcomer does not understand something as s/he tries to make sense of 'what's going on' – the rules, the culture, the norms, s/he will likely ask a few people who are close (by rank, by geography, by departmental affiliation, by situated interaction). If that newcomer has other ideas, philosophies, and thoughts about values, and about how things should work, he/she might reserve those to him- or herself, so as not to lose status. If they are expressed, chances are they will not make a lot of 'waves', as this newcomer is not yet well connected. Therefore, chances that the existing order-of-things will be challenged are low. But newcomers to Wikipedia are slightly different. While status does play a role in Wikipedia, it probably has lesser implications on one's life overall, compared to member status in traditional organizations (where it can have significant effects people's social life and financial situation). And, importantly, whatever is said on the policy discussion page can be seen by the entire community. Thus, simply dismissing someone just because s/he is new, without reasoning, while perhaps possible, is more problematic.

So the process might just start with a new person who lacks knowledge about previous understandings, or an editor who decides to challenge the status quo. In either case, a change process may begin with just one small

question, and an answer. What starts with a cognitive puzzle, turns into an act (asking a question), which really is already an attempted interact with others (who at first may be assumed, or imagined); which then turns into an interact when they answer, and may, to the extent the conversation evolves, become a seed of a structurational process. Observing the online discussions through a collective-sensemaking lens can therefore extend our understanding of how changes to the policy (and then, to the organization) originate and develop from individual attempts to make sense. Through conversation, cognitive efforts turn from individual to social, and commitments are formed and reified in the shape of policy text.

Citing March, Schulz, & Zhou (2000, p. 18), Butler et al. note that *"Because they are explicit and visible... written policies and rules are often sites of conflict" (ibid, p. 2).* They further note that due to these characteristics (visibility, clear boundaries) written policies have greater potential as levers for stakeholders to affect the community, or in other words, to initiate change (ibid, p. 2). Indeed, conflict is evident in many of Wikipedia's policy talk pages which I have reviewed. Yet the discussions I have read reflect different types of conflict – of interest, of world view and of interpretation. Therefore I find the following quote from Barley (1988, p. 51; brought in Orlikowski, 1996, p. 65) more subtle and accurate in describing how policies can be sites in which, and around which, interpretation and negotiation take place: *". . . Because forms of action and interaction are always negotiated and confirmed as actors with different interests and interpretations encounter shifting events (. . .), slippage between institutional templates and the actualities of daily life is probable. In such slippage resides the possibility of social innovation. "*

I would even go further and say this: based on my observations so far, *tensions* between the "organizational templates", i.e. the policies, rules, guidelines and templates of Wikipedia and the "actualities of life" in Wikipedia are not only *probable*, but rather, *constantly present*[9]. The case study of the creation of the "Biographies of Living Persons" policy, as depicted by Forte and Bruckman (2008) provides an example of several such gaps and tensions that formed between Wikipedia's organizational templates and the actualities of life at a certain period. Accumulation of several such tensions was the trigger that drove efforts, acts and interacts of interpretation and sensemaking, which in that particular case initiated an organizational change in the form of the creation of a policy. In other cases, similar efforts often end in reinforcing existing structures.

Conclusion

Wikipedia's discussion pages provide a unique opportunity for micro-level organizational inquiry. These pages, powered by MediaWiki software, have two unique properties (when juxtaposed with discussions that take place in traditional organizations, or with private discussions in other online communication channels): One, they are publically visible, and two, they endure.

---

[9] I wrote this in the original version of this essay. But after thinking about this some more, I tend to think that this is related to the period I have examined, which was a formative period in the life of the community. Given the general pattern of stabilization (, e.g. slowing down of policy making) evident in findings of several works cited here, I believe that the tensions between organizational templates and the "actualities of life" are likely to decline over time (least during "stable" periods), as organizational templates evolve and the organization matures.

For organization members, not only do these pages facilitate conversation – they also accelerate the speed with which people can engage in it, the number of people who can take part in it, and its potential impact (both the immediate impact, and long-term one). For researchers, this allows unparalleled access to huge volumes of organizational discussion.

I was able to track how individual sensemaking efforts turn into interacts, using a mechanism of questions and answers over online discussion pages. People thus engage in a collective sensemaking process. I offer that a prototype of sensemaking - *committed interpretation* (Weick, 1993) can help us understand collective sensemaking processes in Wikipedia, and account for structurational activity that includes the construction of social structures; their ongoing transformation over time; and some trends of stabilization over time.

Interacts between people often become binding, and Wikipedia's discussion environment is especially conducive for creating social commitments, because participation in interaction is volitional, and because discussion pages remain publicly accessible.

While people do not always share meanings and interpretations, they can – and do – achieve temporary, partial interpretations that satisfice their needs, and to which they can commit. Those then become reified on the official policy page, and they stay so, so long as no challenger has managed to convince the majority of those who care that they need to be changed or removed.

This paper makes several contributions to the literatures of organizational studies, computer-supported-collaborative work, and Wikipedia.

To the literature of organizational science (in particular, sensemaking) it contributes an empirical account that grounds the idea of committed interpretation, which I believe to be the first that does so in a radically distributed, open, web-based organization. Weick notes that "We already know that many current ideas about sensemaking assume vertical hierarchies (e.g. uncertainty is absorbed as communications flow upward). What we need to know is what happens to sensemaking when this assumption is replaced by the assumption that structuring unfolds laterally, more like the networks of conversation Winograd and Flores mentioned?" (Weick, 1995, p. 175).

At least in the case I have analyzed, the process model of committed interpretation seems to hold, and I found it helpful for understanding and explaining the phenomena. In fact, as discussed above, the wiki environment amplifies the publicity and irrevocability of volitional interacts, and thus intensifies the process of turning them to commitments. Therefore, it seems plausible that the model would hold for other wiki-based communities. As organizations gradually adopt social software platforms, widely-visible and virtually-permanently accessible communications are likely to become more prevalent, and there is reason therefore to believe that the model presented here will be useful for understanding them as well.

By tracking this collective sensemaking process, I was also able to offer how links are formed between individual and social cognitions, and

provide empirical evidence of the way committed interpretation and collective sensemaking relate to structuration.

To the literature of Wikipedia in particular, and to CSCW in general, this paper contributes a sensemaking perspective on the processes of structuring of Wikipedia's bureaucracy, and a process model of regulation and change, based on the conceptual model of committed-interpretation. This modeling helps us propose explanation for structurational activity that includes several, seemingly unrelated phenomena, including the growth in policy citation counts over time (Beschastnikh et al., 2008), and the process of how social structures (e.g. policies) get enacted and changed within Wikipedia.

I believe this perspective of collective-sensemaking, and the concept of committed-interpretation should prove useful for studying structurational process in other related settings including heavily distributed organizations and online communities.

Acknowledgments

My sincere thanks go to Susan Silbey, Lotte Bailyn and students in their doctoral seminars at MIT who provided constructive feedback in the early stages of this work; to the CSCW coordinator and three anonymous reviewers for detailed and constructive feedback on two earlier revisions of this paper; and to Rob Laubacher, Wanda Orlikowski and Stephanie Woerner for their valuable comments on the final draft.

References

Barley, S. R. (1986). Technology as an occasion for structuring: evidence from observations of CT scanners and the social order of radiology departments. *Administrative Science Quarterly, 31*(1), 78-108.

Barley, S. R. (1988). Technology, Power, and the Social Organization of Work: Towards a Pragmatic Theory of Skilling and Deskilling. *Research in the Sociology of Organizations, 6*, 33-80.

Beschastnikh, I., Kriplean, T., & McDonald, D. W. (2008). *Wikipedian self-governance in action: Motivating the policy lens*. In proceedings of the International Conference on Weblogs and Social Media (ICWSM).

Brennan, S. E. (2005). How Conversation Is Shaped by Visual and Spoken Evidence. In J. C. Trueswell & M. K. Tanenhaus (Eds.), *Approaches to studying world-situated language use: Bridging the Language-as-Product and Language-as-Action traditions* (pp. 95-129). Cambridge, MA: MIT Press.

Bryant, S. L., Forte, A., & Bruckman, A. (2005). *Becoming Wikipedian: transformation of participation in a collaborative online encyclopedia*. In proceedings of the GROUP International Conference on Supporting Group Work, Sanibel Island, FL, USA.

Buriol, L., Castillo, C., Donato, D., Leonardi, S., & Millozzi, S. (2006). *Temporal Analysis of the Wikigraph*. In proceedings of the Web Intelligence Conference

Butler, B., Joyce, E., & Pike, J. (2008). *Don't look now, but we've created a bureaucracy: the nature and roles of policies and rules in wikipedia*. In proceedings of the Twenty-sixth annual SIGCHI conference on Human factors in computing systems (Session: Shared Authoring), Florence, Italy.

Charmaz, K. (2006). *Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis*: Sage.

Czarniawska-Joerges, B. (1992). *Exploring Complex Organizations: A Cultural Perspective*: Sage Publications, Inc.

Forte, A., & Bruckman, A. (2005). *Why do people write for Wikipedia? Incentives to contribute to open*. In proceedings of the GROUP 05 workshop: Sustaining community: The role and design of incentive mechanisms in online systems. , Sanibel Island, FL, USA.

Forte, A., & Bruckman, A. (2008). *Scaling consensus: increasing decentralization in Wikipedia governance*. In proceedings of the Hawaiian International Conference of Systems Sciences (HICSS), Hawaii.

Forte, A., Larco, V., & Bruckman, A. (2009). Decentralization in Wikipedia Governance. *Journal of Management Information Systems, 26*(1), 49-72.

Kellogg, K. C. (Forthcoming). Operating Room: Relational Spaces and Micro-Institutional Change in Surgery. *American Journal of Sociology, 15*(3).

Kittur, A., Suh, B., Pendleton, B. A., & Chi, E. H. (2007). *He says, she says: Conflict and coordination in Wikipedia*. In proceedings of the Twenty fifth SIGCHI conference on Human factors in computing systems (CHI '07).

Kriplean, T., Beschastnikh, I., McDonald, D. W., & Golder, S. A. (2007). *Community, Consensus, Coercion, Control: CS*W or How Policy Mediates Mass Participation*. In proceedings of the Conference on Supporting Group Work.

Malone, T. W. (2004). *The Future of Work: How the New Order of Business Will Shape Your Organization, Your Management Style, and Your Life*. Boston, MA: Harvard Business School Press.

Malone, T. W., Laubacher, R. J., & Scott Morton, M. S. (2003). *Inventing the Organizations of the 21st Century*: The MIT Press.

March, J. G. (1981). Footnotes to organizational change. *Administrative Science Quarterly, 26*(4), 563-577.

March, J. G., Schulz, M., & Zhou, X. (2000). *The dynamics of rules: Change in written organizational codes*: Stanford University Press.

Morgan, J. T., & Zachry, M. (2010). *Negotiating with angry mastodons: the wikipedia policy environment as genre ecology*. In proceedings of the 16th ACM international conference on Supporting group work (GROUP '10), New York, NY, USA.

Orlikowski, W. J. (1996). Improvising Organizational Transformation Over Time: A Situated Change Perspective. *Information Systems Research, 7*(1), 63-92.

Reagle, J. (2005). Is the Wikipedia Neutral. Retrieved May 13, 2010, from http://reagle.org/joseph/2005/06/neutrality.html

Suh, B., Chi, E. H., Kittur, A., & Pendleton, B. A. (2008). *Lifting the veil: improving accountability and social transparency in Wikipedia with wikidashboard.* In proceedings of the Twenty-sixth annual SIGCHI conference on Human factors in computing systems (CHI '08).

Viégas, F., Wattenberg, M., & McKeon, M. (2007, July 22-27). *The hidden order of Wikipedia.* In proceedings of the OCSC 2007 - Online Communities and Social Computing: second international conference, held as part of HCI International, Beijing, China.

Wales, J. (2005, July). Jimmy Wales on the birth of Wikipedia. Retrieved May 17, 2010, from http://www.ted.com/talks/lang/eng/jimmy_wales_on_the_birth_of_wikipedia.html

Weick, K. E. (1993). Sensemaking in organizations: Small structures with large consequences. In J. K. Murnighan (Ed.), *Social psychology in organizations: Advances in theory and research* (pp. 10-37): Prentice Hall College Division.

Weick, K. E. (1995). *Sensemaking in organizations*: Sage.

Essay 2

Using Collective-Intelligence:

Combining Human and Machine Predictions in Semi-Structured Environments

Yiftach Nagar

Thomas W. Malone

Abstract

Prior research has shown that in relatively structured, stable environments, statistical models are almost always at least as good as human experts at making predictions, and often substantially better. However, many important prediction problems in the real world arise in *semi-structured environments* where data are difficult to codify or quantify, where patterns are difficult to discern, and where changes occur unexpectedly. We hypothesize that in these environments, where it may be difficult – or even impossible – to build reliable and dependable predictive models, combining people and computers can lead to improved predictions. To test this hypothesis, we conducted two laboratory experiments in which we used prediction markets, human judgment, and averaging to combine predictions from groups of people and artificial intelligence agents. We found that the combined predictions were both more accurate and more robust than predictions made by groups of only people or only machines. We discuss the appropriateness of these methods in different contexts.

Introduction

Substantial evidence from multiple domains suggests that predictive models usually yield better (and almost never worse) predictions than do individual human experts (e.g. Dawes, Faust, & Meehl, 1989; Dawes & Kagan, 1988; Grove, Zald, Lebow, Snitz, & Nelson, 2000). Whereas models (or machines) are better at information processing and are consistent (Einhorn, 1972; Goldberg, 1970), we humans suffer cognitive and other biases that make us poor judges of probabilities (c.f. Kahneman, Slovic, & Tversky, 1982; Kahneman & Tversky, 1973; Lichtenstein, Baruch, & Phillips, 1982; Rabin, 1996). In addition, *"Such factors as fatigue, recent experience, or seemingly minor changes in the ordering of information or in the conceptualization of the case or task can produce random fluctuations in [human] judgment"* (Dawes et al., 1989). Human *groups* also often exhibit phenomena such as *groupthink* (Janis, 1972; Janis & Mann, 1977) and *group polarization* (Brown, 1986, pp. 200-248; Myers & Lamm, 1976) that negatively affect their judgment and their ability to make reliable predictions.

Nevertheless, human intelligence and judgments are often still valuable in predicting events in real-life situations, as we still possess three critical elements: first, human intelligence still outperforms even the most advanced computational systems when it comes to the acquisition and understanding of many kinds of information. This is especially true for unstructured information (Einhorn, 1972; Kleinmuntz, 1990), and this human advantage– while eroding surprisingly quickly in some domains – is not likely to completely disappear anytime soon. Second, humans possess abstract knowledge about the world, which is still far too expansive to be included in models, and which enables us to

learn and make inferences on the basis of limited data (Negash, 2004; Tenenbaum, Kemp, Griffiths, & Goodman, 2011). Third, this knowledge of the world also enables us to easily identify "broken-leg" situations (Camerer & Johnson, 1991; Johnson, 1988; Meehl, 1954), in which the rules normally characterizing the phenomenon of interest do not hold, and to interpret their potential consequences. As Téglás et al. (2011) summarize: *"[The] ability to flexibly combine multiple sources of information and knowledge to predict how a complex situation will unfold is at the core of human intelligence and is one of the biggest missing links in building artificial intelligence systems with humanlike 'common sense'"*.

Several researchers (e.g. Blattberg & Hoch, 1990; Bunn & Wright, 1991; Einhorn, 1972; Seifert & Hadida, 2013) considered these different strengths of human experts and models, and contemplated the complementary nature of humans and models in making predictions. For instance, in Blattberg and Hoch's study, a "50% Model + 50% Manager" heuristic improved forecast quality of catalog fashion sales, and coupon redemption rates. However, there has been relatively little empirical work along these lines.

It is also worth noticing that previous discussion on combining human and model predictions does not stress the potential of improving predictions by combining predictions from *multiple* humans and models. A vast body of theoretical and empirical research suggests that combining forecasts from multiple independent forecasters that have relevant

knowledge and information leads to increased forecast accuracy[10]. This virtually unanimous result holds whether the forecasts are based on human judgment or mathematical models (Armstrong, 2001; Clemen, 1989). Further, because it may be difficult or impossible to identify a single forecasting method that is the best (Makridakis & Winkler, 1983), *combining forecasts is less risky in practice* (Hibon & Evgeniou, 2005). Research on pattern-recognition classifiers in artificial-intelligence offers similar conclusions (cf. Duin & Tax, 2000; Ho, Hull, & Srihari, 1994; Kittler, Hatef, Duin, & Matas, 1998; Lam, 2000; Suen & Lam, 2000).

Weaving together these threads of inquiry, it seems plausible that combining predictions from multiple humans and machines could potentially emphasize their relative advantages, mitigate their respective flaws, and, thus, yield better predictions than those created by either humans or machines alone. Particularly, here we offer that this approach can be beneficial for a class of important real-world situations that might be called *semi-structured environments* – where patterns are difficult to discern, where data are difficult to codify and/or quantify, or where sudden changes might occur unexpectedly. For instance, in complex situations involving the actions of human groups in volatile environments (e.g. political parties, regulators, or customers and business competitors in a fashion-driven business), there is often a fair amount of relevant data that can be analyzed, but it may be difficult to find patterns and formulate all the rules governing the phenomena of interest, and some important

---

[10] (For discussion of the philosophical and the mathematical principles underlying the logic of combining forecasts, see Armstrong, 2001; Larrick & Soll, 2006; Makridakis, 1989; Sunstein, 2005, pp. 972-974; Winkler, 1989).

factors may be very difficult to quantify at all. In some contexts (e.g. military) that involve tactics and strategy, parties might even engage in deception to hinder prediction of their moves. In such domains, machine learning and other quantitative methods can be useful in building sophisticated and adaptive models based on potentially vast amounts of data (for recent examples, see Bohorquez, Gourley, Dixon, Spagat, & Johnson, 2009; Mannes et al., 2008); and humans' tacit knowledge, ability to acquire unstructured information, and intuition can help in both information acquisition, and in identifying "broken-leg" situations and preventing catastrophic prediction errors.

Methods for Combining Predictions

But how to combine? Many different ways of combining predictions are explored in the literatures of forecasting and model fusion, including simple and weighted averaging, median, majority voting, etc., as well as techniques that involve learning, e.g. Bayesian learning. Both theoretical and empirical comparisons have shown that no single method or rule of combination is best under all circumstances (see, for example, Armstrong, 1989; Clemen, 1989; Duin & Tax, 2000; Kuncheva, 2002; Lam, 2000; Winkler, 1989).

As our primary interest is in combining predictions from humans and agents, we investigate three different combination mechanisms (two mechanical, one judgmental), focused not on comparing those mechanisms to each other, but rather on the question of whether the combined human-agent predictions are better than those of humans, or agents. The methods we chose are *simple average, prediction markets*, and *human judgment.*

The simple average yields predictions that are at least as accurate as those of the average forecaster (Larrick & Soll, 2006); usually more accurate than most individual forecasts (Stewart, 2001); and, in many cases, better than those of the best individual forecaster (Larrick & Soll, 2006). Compared with other combining mechanisms, the simple average is the most robust against 'classifier peculiarities' (Diebold & Mariano, 1995; Lam, 2000). Accordingly, it is often advised as a good default mechanism for combining predictions (Armstrong, 2001), and thus, a natural candidate for our experiments.

Prediction markets[11] are a more novel idea. Over the past decade, following their success in public settings (Berg, Forsythe, Nelson, & Rietz, 2008), many companies started using them to efficiently aggregate predictions from employees (Cowgill, Wolfers, & Zitzewitz, 2008; Malone, 2004; Sunstein, 2006; Surowiecki, 2004; Wolfers & Zitzewitz, 2004). Empirical investigations have shown that indeed they yield predictions that are usually at least as accurate and as calibrated as other methods traditionally used for forecasting (K.-Y. Chen & Plott, 2002; Cowgill et al., 2008; Hopman, 2007; Ortner, 1997; Spann & Skiera, 2009). Prediction markets also fare well against other methods of combining predictions such as simple average, weighted average and logarithmic regression (Berg, Nelson, & Rietz, 2008; Y. Chen, Chu, Mullen, & Pennock, 2005). Notably, in a two-year test on geopolitical questions, the DAGGRE

[11] Also known as information markets, decision markets, electronic markets, virtual markets, idea futures, event futures and idea markets (Tziralis & Tatsiopoulos, 2007; Wolfers & Zitzewitz, 2004)

prediction market performed 40% better than the simple average (Twardy et al., 2014)

While prediction markets have mostly been used to aggregate predictions from humans, there is no reason why the mechanism cannot also be used to aggregate predictions from software agents; yet this option remains mostly unexplored. One exception is a study by Perols, Chari, and Agrawal (2009) who used a prediction market to combine predictions from machine classifiers. In their experiment, depending on the setting, the market mechanism either outperformed or performed on par with 3 benchmark combination mechanisms: simple average, weighted average, and majority. We are not aware, however, of any previous work using prediction markets to combine human and model predictions.

It is certainly possible that, in some scenarios, prediction markets will provide only minute improvements in accuracy over other methods, as two recent studies (Goel, Reeves, Watts, & Pennock, 2010; Perols et al., 2009) suggest; and costs of implementation, set-up and training should be considered. However, prediction markets may be appealing in some settings for reasons beyond accuracy improvement. First, they incentivize participation of precisely the people (or agents) who can contribute the most. There is little incentive for people (and agents) whose predictions do not increase the accuracy of the market to participate because they are unlikely to profit from participation. For instance, in routine situations where automated agents are doing a good job of predicting, humans have little reason to intervene. But in unusual situations where humans can see that the agents' predictions are very wrong, the humans have a clear incentive to intervene. Also, by giving people incentives to design and run their own 'pet' agents, this framework potentially incorporates continuous improvement into the forecasting process.

In addition, by increasing the game-like challenge of participation and by tying compensation to performance, prediction markets can increase both extrinsic and intrinsic motivations for people to participate. For instance, human participants in prediction markets have an economic incentive to gather more information and implicitly share this information by using it in the market (Cowgill et al., 2008; Hayek, 1945). The sense of wide participation may also increase the legitimacy and acceptance of the predictions made.

Finally, as Perols et al. (2009) note, unlike some combination methods that require learning and setup, prediction markets can adjust to changes in base-classifier composition and performance without requiring offline training data or a static ensemble composition configuration. For all these reasons, prediction markets are a potentially useful and appealing mechanism for dynamically combining predictions from a varying population of humans and agents in real organizational settings.

Finally, we consider one other combination method: human judgment. While some researchers argued that mechanical combination might be better, in many real-life organizational settings, spanning medical, business and military contexts, humans are given the final word, for reasons that may go beyond pure considerations of forecast accuracy, including legal liability and acceptance by target audiences. We therefore also thought it worthwhile to inquire whether humans informed by model predictions will make forecasts that are better. While human forecasters excel at identifying and interpreting "broken-leg" cues, they also tend to overweigh them and miss or under weigh latent regularities (Johnson, 1988). Being informed with the model predictions, while being aware of what the model might miss, may allow experts to adjust and improve predictions.

It is important to realize that we are not hypothesizing that combining human and machine predictions in any of these manners is *always* better, only that it is *sometimes* better. As such, our results can be seen as an existence proof of one situation in which it is better. In the conclusion of the paper below, we speculate about the range of other situations in which this approach may be superior.

Overview of Experiments and Results

To test our hypotheses, we performed two studies, in which predictions from both groups of people and groups of artificial intelligence agents were combined. In the first study, we used prediction markets to combine predictions, and we compared three conditions: markets with only human participants, markets with only artificial-intelligence agents, and markets where humans and artificial-intelligence agents participated simultaneously. In the second study, we tested two additional mechanisms of combining human and agent predictions: (a) simply averaging human and agent predictions, and (b) letting humans adjust their own predictions after seeing the agents' predictions.

We used several different criteria to compare the quality of predictions, including multiple measures of accuracy, calibration and discrimination, and risk-to-reward measures. Regardless of combination mechanism, combined predictions from humans and AI agents were overall better than predictions created by combining human-only, or agent-only predictions.

In the following sections we discuss the studies and the results in detail.

Study 1

Method

Subject matter for predictions: Participants were shown videos of an American football game and, after each play, were asked to predict whether the next play would be a run or a pass. This domain was chosen, in part, as an analog for a wide range of other domains including, for instance, the next actions of terrorist groups, military enemies, or business competitors. Importantly, this setting also enabled us to emulate a realistic situation, as humans had access to a wider range of unstructured information than the agents did (e.g., all the images and commentary in the football game video).

Human Participants: Participants were recruited from the general public via web advertising. We encouraged the participation of football fans by stressing the fun of watching football and by stating that knowledge of football could help make higher profits, but we did not require specific knowledge about football. Compensation to participants included a base payment and an additional performance-based bonus proportional to the ending balance in each participant's prediction market trading account. Bonuses could reach up to 75% of the base pay.

AI agents: We used standard 3-layer artificial neural-net agents developed using the *JOONE* open-source package[12]. For each play, the agents had three pieces of previously coded information: the down

[12] Available at http://sourceforge.net/projects/joone/

number, the number of yards to first down, and whether the previous play was a run or pass. The agents were all trained on one set of plays from a previous game. In addition, the agents considered the market price and traded only if they were confident about their prediction. Of course, there are many other possible kinds of artificial intelligence approaches that could be used here, many of which could presumably make the agents more accurate. As noted above, however, our goal was not to create the best possible artificial intelligence agents, merely to create one example of such agents for experimentation.

Prediction markets: We ran the experiments using a custom-tailored version of the *ZOCALO* open-source prediction markets platform[13], and employed its automated market maker to simplify trading and ensure liquidity in the markets. Each human participant sat at a separate computer from which they could access the prediction market software. Participants could not see each other's screens.

Procedure: We conducted 20 laboratory sessions, in each of which groups of 15 – 19 human subjects participated in prediction markets, both with and without computer agents (median group size was 18; mean 17.55; mode 19; totaling 351 participants overall). Subjects first completed a short questionnaire where they reported their level of interest in football and their self-assessed level of knowledge of the game. They also answered a 20-question quiz about football, designed to estimate their level of expertise more objectively. These questionnaires were used as controls rather than a part of our main inquiry.

---

[13] Available at http://zocalo.sourceforge.net/

After initial explanation[14] and training rounds, each experimental session included 20 plays. The same set of 20 plays[15] was shown in all the sessions. For each play, a short video excerpt from the game was shown to all participants. The video was automatically stopped just before the team possessing the ball was about to start a play. At that stage, a new online prediction market was opened and the group of participants (either human participants only, or human participants along with AI agents[16]) started trading contracts of RUN and PASS. (All plays that were not RUN or PASS were eliminated from the video). The market was closed after 3.5 minutes, and the video continued, revealing what had actually happened, and stopping before the next play.

The AI agents participated either in the first half (first 10 plays) or the second half (last 10 plays) of the experiment (according to a random draw previously performed). Human participants were told that AI agents would trade in some of the markets but were not told in which, and could not generally tell. Thus in each lab session we collected data from 10 'human only' markets and 10 'hybrid' (humans and agents) markets. In addition we ran 10 "computer-only" experimental sessions with no human

---

[14] Participants were given an elaborate verbal explanation on the goal of the experiment, and on trading in the prediction market. In addition, participants were prompted to read a short manual the day before coming to the lab, doing which – as they were truthfully told, would raise their chance to succeed in the markets and make a higher bonus. We regularly checked by show of hands how many of them actually read the manual and the overwhelming majority did. The manual was also available on participants' screens, though they rarely, if ever, referred to it during the sessions.

[15] Taken from the 2008 Fiesta Bowl game between West Virginia University and Oklahoma University.

[16] We ran 10 neural-net agents. They used the same code and same training dataset, but their logic of trading was also based on the market price, and they were started in a staggered manner so that they encountered different market conditions.

participants, where the agents traded in the markets for all 20 plays. We thus got a total of 600 observations (10 observations of each of our 3 conditions for each of 20 plays).

In our analysis, we took the market closing price as representing the collective group estimation of the probability of the football team to either RUN or PASS the ball (see Wolfers & Zitzewitz, 2004, 2006).

Results

Prediction quality is a multidimensional concept that aims to capture the degree of correspondence between predictions and observations. There are many measures by which predictions can be assessed, but no single measure is sufficient for judging and comparing forecast quality (Jolliffe & Stephenson, 2003). Thus, assessment of prediction quality is a matter of analyzing and understanding trade-offs. To compare the three groups of predictors, we therefore look first at three criteria common in the forecasting literature: ***Accuracy, Reliability (a.k.a Calibration)*** and ***Discrimination*** which, combined, help understand those trade-offs. We augment this analysis with a comparison of accuracy vs. variability, using the Sharpe ratio (Sharpe, 1966, 1994), commonly used in finance to compare reward-vs.-risk performance. Finally, we also present an analysis based on the ***Receiver-Operating-Characteristic (ROC)*** approach (Swets, 1988; Swets & Pickett, 1982; Zweig & Campbell, 1993) that has been established and widely accepted in many domains as a method of assessing and comparing predictors who make predictions about binary events.

Accuracy

Accuracy is a measure or function of the average distance/error between forecasts and observations. A common way to assess the accuracy of predictions and to compare the skill of the people or methods that created them is to use a scoring rule. Table 1 summarizes the evaluations of accuracy for the humans-only markets, agents-only markets, and hybrid markets, over the experimental play set, according to three popular scoring rules: the Mean Absolute Error (MAE), the Mean Square Error (MSE[17]), also known as the Brier Score (Brier, 1950), and the Log Scoring Rule[18] (LSR, introduced by Good, 1952). The lines to the right of the MAE and the LSR scores indicate where the differences between the scores of the different predictors were found statistically significant[19] ($p<0.05$). Under all of these scoring rules, a score that is closer to zero is better, and under all of them a perfect predictor who assigns a probability estimation of 100% to actual events and a probability of zero to all other potential options will score zero.

---

[17] Several authors have noted limitations of comparing forecasts/forecasters using the Brier score (cf. Clements & Hendry, 1993; Diebold & Mariano, 1995; Ferro, 2007; Jewson, 2004). Nevertheless it is one of the most commonly used measures in the forecasting literature.

[18] To keep the logic of other rules we reversed the original Log scores, such that a lower score is better. $LSR = 2\text{-}Log_{10}(P)$, where P is the prediction (market closing price) of the actual outcome.

[19] To compare the conditions we built a mixed model to account for nesting, and used SAS's PROC MIXED (Littell, Milliken, Stroup, & Wolfinger, 2006) with the first-order autoregressive AR(1) error-covariance-matrix structure (ibid., pp. 175-176). The squared errors are not normally distributed, which hinders a parametric statistical comparison of the MSE scores. Therefore significance tests for this column are not shown. Distributions of the absolute errors and of the log-predictions are quasi-normal. Variability of group level aggregates of the level of interest in football, or knowledge of football had no significant effect on the results.

Table 1 – Accuracy of Prediction Markets

Market Type	Scoring Rule: Mean Absolute Error	Mean Squared Error	Log Scoring Rule (LSR)
Humans-only	0.42	0.20	0.25
Agent-only	0.35	0.17	0.23
Hybrid	0.35	0.15	0.21

Albeit popular, these comparisons should be interpreted with care.. Scoring rules (to be exact: *proper* scoring rules[20]) are useful in eliciting honest probability estimations, but, as Winkler (1969) cleverly points out, using them to evaluate and rank predictors ex post may be misleading, as it confounds the measurement of accuracy with the cost function of errors. Different scoring rules punish small and large errors to different extents, and can yield contradicting results when used to rank predictors. Indeed in our table, the Hybrid markets are more accurate, on average, than the Agent-only markets according to the MSE and the LSR, but not according to the MAE where they tie.

It is up to the decision maker, therefore, to select the rule to be used for evaluation, and this should be done according to the nature of the setting and the corresponding cost functions. For example, in weather forecasting, small errors are tolerable on a daily basis (say, ±1 degree in temperature predictions), but big errors (predicting a very hot day which

[20] A scoring rule is proper if the forecaster maximizes the expected score for an observation drawn from the distribution F if he or she issues the probabilistic forecast F, rather than $G \neq F$. Practically speaking this means that proper scoring rules encourage the forecaster to make careful assessments and to honestly report his or her true beliefs (Gneiting & Raftery, 2007)

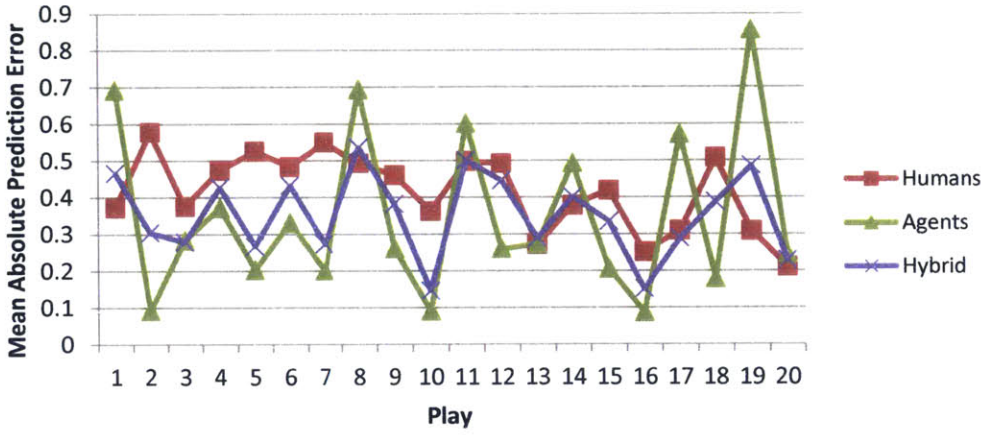
turns out to be very cold, or failing to predict a tornado) are not. In an industrial mass-production setting, on the other hand, it may be OK to throw away a unit due to a large prediction error on rare occasions, but precision is very important on a regular basis. While there may be some ambiguity in selecting a scoring rule when the cost of errors is unknown, in our case it appears that the number of large errors matters more than the average accuracy (e.g. it is likely that a prediction of 90% and prediction of 95% for a PASS attempt by the offense team would both translate to the same decision by the defense team) and hence, the MSE and the LSR seem more appropriate than the MAE.

Taken together, these results suggest that the hybrid markets were the most accurate. We also note that although the agents were very simple, on average the agent-only markets were more accurate than the human-only markets, as one might expect based on previous evidence in the literature.

A deeper look at the play level

A deeper look at the play level provides better understanding of the behavior of the predictors and reveals several interesting patterns. Figure 1 depicts the mean absolute prediction error (average of 10 observations from 10 markets) of each condition, per play.

Figure 1 – Mean Prediction Errors of Human, Agents and Hybrid Markets



We note a strong interaction between condition and play. As could be expected, humans and agents predicted differently on different plays. While *on average* agents were more accurate than humans (i.e. had smaller errors), in several cases they made severe errors while humans predicted correctly (conspicuously: plays 1, 8, 19). But why? Informal interviews with participants indicated that as we hypothesized they incorporated unstructured information (which was not readily available to the agents) into their decision-making. For example, some participants reported that commentary by anchors was helpful, and several others mentioned that the body language of players was revealing. Notably, some of the participants we talked with said they gleaned from the video the formation of the offensive and the defensive teams. For example: before both play 1 and play 29, the team playing offense formed a "Shotgun"

formation[21], with a running-back standing next to the quarterback, which to football savvy fans implies a higher probability for a pass attempt. In both those plays, the 'human-only' markets clearly indicated a pass (70% and 77% on average) whereas the 'all-agents' markets indicated a RUN (31% and 14.5% predictions for PASS). As hypothesized, in these cases the "broken-leg" cue helped humans make the correct prediction where the agents were wrong, and they were able to influence the combined prediction through the market mechanism.

Beyond Mean Errors: Considering Prediction-Error Variability

Measures of accuracy alone do not provide sufficient information to convey the complexity of the data, as they are essentially comparisons of single numbers representing entire distributions. For instance, if two prediction methods have the same expected average error, but one method has much greater variability in error, the one with less variability would usually be preferred. The same is true in finance, as well: Of two investments with equal expected value, the one with less risk would generally be preferred. As another way of evaluating prediction methods, therefore, we use a method originally developed to measure the *reward-to-risk* performance of financial investments. After assigning economic values to the predictions using scoring rules, we use the ex post Sharpe ratio (Sharpe, 1966, 1994), to consider accuracy against variability of prediction errors.

---

[21] Mallory, B., & Nehlen, D. (2006, ch. 7-8). *Football offenses & plays*: Human Kinetics Publishers

To keep with the familiar logic of the Sharpe ratio that assumes a higher positive financial return is better, we adjust our scoring rules such that the adjusted MAE score (AMAE) equals 1-MAE and the adjusted MSE score (AMSE) equals 1-MSE. The adjusted Log score is log10(P) where P is the prediction (market closing price) of the actual outcome. We calculated the Sharpe ratio according to equations 3-6 in Sharpe (1994, p. 50). As a simple and straightforward benchmark, we use an "ignorant" predictor who bets 50% PASS all the time (and whose error variance is therefore zero). The corresponding AMAE, AMSE and ALSR for the benchmark predictor are therefore 0.5, 0.75 and 1.699, correspondingly. The results are summarized in Table 2.

Table 2 - Ex Post Sharpe Ratio for Prediction Markets, Under 3 Scoring Rules

	Scoring Rule		
	AMAE (Benchmark = 0.5)	AMSE (Benchmark = 0.75)	ALSR (Benchmark = 1.699)
Humans-only	0.54	0.40	0.41
Agents-only	0.67	0.39	0.37
Hybrid	0.91	0.73	0.72

Clearly, the hybrid markets yield the highest Sharpe ratio and outperform both the human-only and agent-only markets. This result holds under all three scoring rules. According to the Sharpe ratio index criterion, therefore, the Hybrid markets are more robust, offering a better trade-off between prediction accuracy and variability.

Calibration and Discrimination

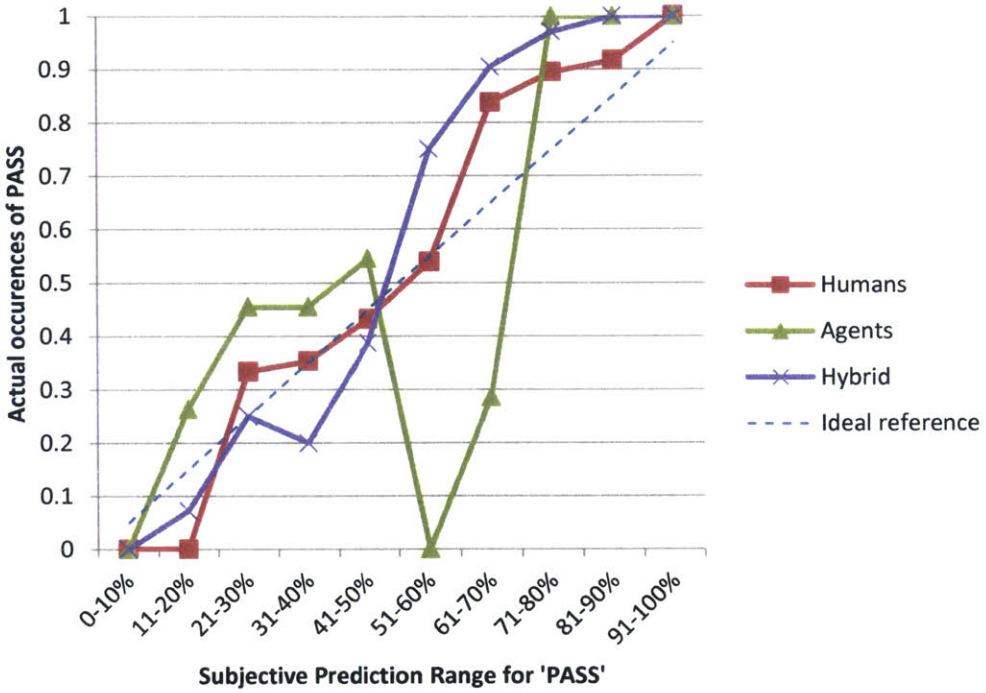
Reliability (Murphy & Winkler, 1977), (also: Calibration, e.g. Lichtenstein et al., 1982), refers to the degree of correspondence between

forecast probabilities and actual (observed) relative event frequencies. For a predictor to be perfectly calibrated, assessed probability should equal percentage correct where repetitive assessments are being used (ibid.).

Calibration diagrams, built by binning predicted probabilities into 10% bins, are commonly used to portray observed event frequencies against predicted probabilities. In Figure 2, we depict the calibration diagram for our 3 conditions. The dotted straight diagonal line stretching from (0,0) to (100,100) represents the ideal reference of a hypothetical perfectly-calibrated predictor. Evidently, both the human and hybrid markets are reasonably calibrated, while the agents are not.

Discrimination (a.k.a Resolution) taps forecasters' ability to do better than a simple predict-the-base-rate strategy. Observers get perfect discrimination scores when they infallibly assign probabilities of 1.0 to things that happen and probabilities of zero to things that do not (Tetlock, 2005, pp. 47-48, 274). It is important to note that calibration skill and discrimination skill are not identical. For example, a predictor that always predicts the base-rate of the event will be perfectly calibrated but will score very low on discrimination (for such a predictor, the calibration plot will only include a single point, on the diagonal reference line).

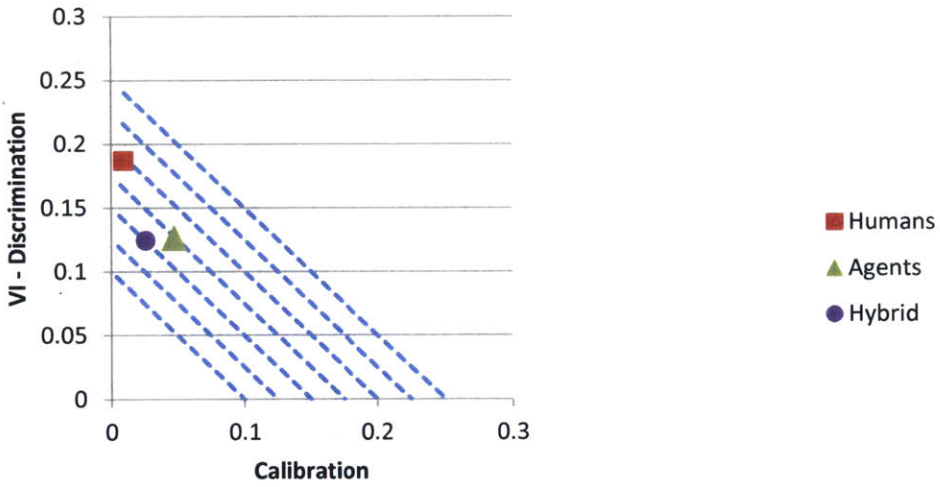
Figure 2 – Calibration of Prediction Markets



It has been offered that the MSE can be decomposed as VI+CI-DI where VI is the variability index representing the uncertainty of the phenomena, CI is the calibration index of the forecasts and DI is the discrimination index of the forecasts (Murphy, 1973; Murphy & Winkler, 1987; see also Tetlock, 2005, pp. 274-275). While the MSE may have drawbacks as a criterion by which to judge the quality of predictions, this decomposition seems nevertheless useful in orienting our understanding of the trade-off between calibration and discrimination of our predictors. Given that the variability of the events in our case is identical (VI=0.24) for the 3 conditions we want to compare (since they made predictions about the same events), we can draw a plot of (Variability – Discrimination) vs. Calibration for each predictor. For a given variability, we can also draw "efficient front" isopleths of MSE.

We present such a plot depicting the performance of our 3 conditions in Figure 3. In this plot, the more calibrated a predictor is, the more to the left it would appear (CI closer to zero is better). The more discriminating a predictor is, the lower it would appear. It is evident in this plot that the Hybrid markets were about as discriminating as the agent markets, but more calibrated. It is also clear that compared to human markets, the hybrid markets were slightly less calibrated, but more discriminating. Overall, the hybrid markets are on a more efficient front compared to both agents markets and human markets – as reflected in the MSE scores.

Figure 3 – [Variability – Discrimination] vs. Calibration (with MSE Isolines. VI=0.24)



ROC Analysis

Our comparisons of accuracy, and of the Sharpe ratio, both rely on attaching costs to prediction errors using scoring rules. While we used common rules, they may not represent the actual economic value of predictions (or corresponding errors), and in reality, it is not always

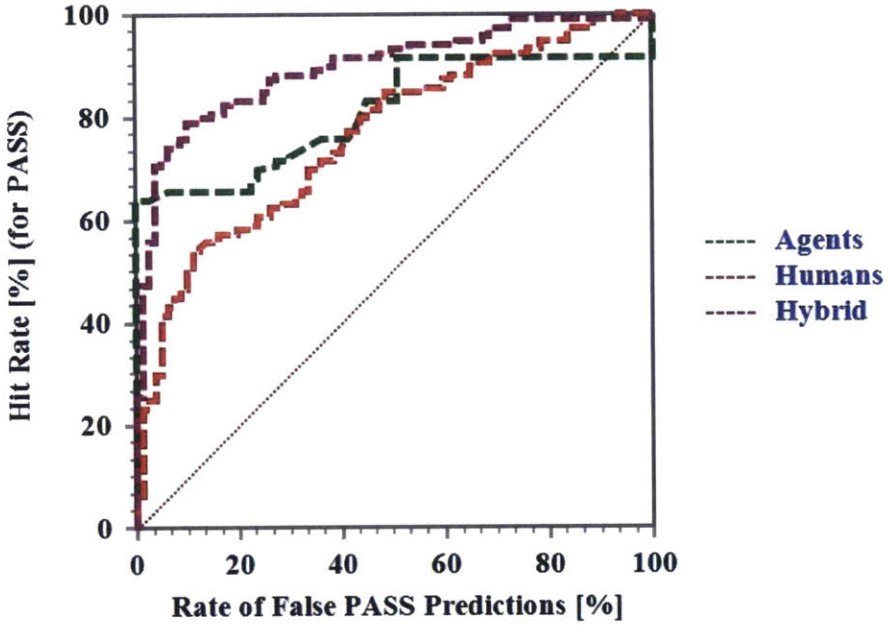
possible to determine the true cost of prediction errors. Abdellaoui, Bleichrodt, l'Haridon, and Paraschiv (2013) further argue that these costs are volatile and may be subject to framing.

The Receiver-Operating-Characteristic (ROC) is an established methodology for evaluating and comparing the performance of diagnostic and prediction systems, which does not rely on their unknown economic value, and hence, can provide additional support for our conclusions. ROC has been widely used in many different domains including signal detection, radiology, weather forecasting, psychology, information retrieval etc. (Swets, 1973, 1988; Swets & Pickett, 1982; Zweig & Campbell, 1993). ROC curves are obtained by plotting the hit rate (i.e. correctly identified events) versus the false alarm rate (incorrect event predictions) over a range of different thresholds that are used to convert probabilistic forecasts of binary events into deterministic binary forecasts (Jolliffe & Stephenson, 2003, p. 211). The ROC, plotted for a range of different thresholds, offers a more credible view of the entire spectrum of accuracy of the different predictors (Zweig & Campbell, 1993, pp. 563-564), and serves to highlight the tradeoff between sensitivity and specificity of each predictor. The area under the curve (AUC) serves as a measure of the quality of the predictions, with a perfect predictor scoring 1. The ROC curves[22] of our conditions are presented in Figure 4, and the areas under the curves are depicted in Table 3.

---

[22] The construction of ROC curves assumes a reference category. In this case we took "PASS" to be the event of interest and thus, a hit is the correct prediction of a PASS, and a false positive is the prediction of a PASS that turns out to be a RUN. Had we used

Figure 4 - ROC Curves for predictions of football plays by Human-only, Agent-only and Hybrid prediction markets
(20 plays, 10 observations of each play by each condition)



RUN as the event of interest we would have gotten transformed curves, but they would be equivalent, and the area under the curves would be the same as in the plot we display.

Table 3 – Comparison of the Areas Under the ROC Curves

Forecasters	Area under ROC Curve[23]	SE[24]
Humans-only	0.76	0.033
Agents-only	0.81	0.031
Hybrid	0.90	0.022

This result suggests that the hybrid prediction markets may provide a better trade-off between sensitivity and specificity when compared to either humans-only or agents-only prediction markets. In that, it echoes our previous analyses.

[23] The areas under the curves were calculated in MedCalc software (Schoonjans, Zalata, Depuydt, & Comhaire, 1995). MedCalc is available from http://www.medcalc.be/

[24] There is no widely accepted way to test the statistical significance of differences of areas under the ROC curve for repeated measurements of the same events. We calculated standard errors using the method offered by DeLong, DeLong, & Clarke-Pearson (1988).

Study 2

Study 1 showed that indeed, prediction markets can successfully be used to aggregate and combine predictions from humans and agents and that those hybrid markets can yield higher quality, more robust predictions than do markets of humans, or agents. As discussed above, we wanted to explore additional ways of achieving synergy between humans and agents for improving predictions, and in study 2 we therefore tested two other ways of combining their predictions. First, we tried to simply average human and agent predictions. Second, we wanted to check whether explicitly informing people of the agents' prediction and letting them judge this input may help them improve their own predictions. Eliminating the need to run prediction markets also provided us more time during the sessions, which enabled us to broaden our tests to a larger set of 55 plays.

Method

We recruited 203 participants, from the general public, mainly through web advertisements as we did for study 1. We held sessions with varying numbers of participants, as there was no meaningful group interaction during the study. As with study 1, we encouraged the participation of football fans but did not require participants to be experts. The compensation scheme was similar to that of study 1, except for the bonus, which was proportional to the mean squared error of the subject's predictions during the entire session.

Participants sat in front of personal computers, and worked independently, using headphones and dividers so that they could not see or hear the other participants' predictions. Following initial

administration, explanations and several training rounds, participants were asked to view video excerpts from the same football game we used in study 1, and provide their predictions in the form of probability estimations (a number between 0 and 100) that the playing team would pass the ball. After providing an initial estimation, each subject was informed what the agents' prediction was for the play she or he just predicted and was asked to provide a second, revised prediction, considering the agents' input (which we collected earlier). After that, in a similar manner to study 1, the subject viewed the next video clip which revealed what the actual play was, and stopped before the next play. We used online questionnaire software to conduct the experiment, and configured it such that participants could not make a prediction before seeing the entire clip, and could not go back and change a prediction once it was committed.

We collected 3 datasets: agent predictions (average of 10 runs), the humans' initial predictions, and humans' second predictions that were informed by the agents' predictions. We also computed a fourth dataset by averaging of the agents' predictions with the initial predictions from the human participants. To make the analysis similar to study 1, we randomly divided our participants into 12 groups, ranging from 15 to 19 participants in each (median: 17, mean: 16.9). In order to make sure our statistical analyses were valid we repeated them with 3 different random groupings, and the results we show here are the average of these 3 groupings. We only report a result as significant if the p-value was <0.05 in all 3 analyses.

Results

Accuracy

Table 4 summarizes the evaluations of accuracy for the four conditions over the experimental play set, according to the MAE, MSE and LSR. The lines to the right of the MAE and the LSR scores indicate where the differences between the scores of the different predictors were found statistically significant ($p<0.05$). The mean squared errors were not normally distributed, hindering statistical comparison of the means.

Table 4 – Accuracy Results: Study 2

	Scoring Rule		
	Mean Absolute Error	Mean Squared Error	LSR
Agents	0.40	0.26	0.34
Uninformed Humans	0.44	0.21	0.27
Avg. of Humans & Agents	0.42	0.21	0.26
Informed Humans	0.41	0.20	0.25

Like the hybrid markets in study 1, informed humans were more accurate than the agents or the uninformed humans under the MSE and the LSR. Under the MAE, agents' predictions were more accurate. This result implies that, as in study 1, while *on average* the agents' predictions were more accurate, they had a higher number of large errors. We check this in the following two sections. Interestingly, informed humans' predictions were more accurate, on average, than the simple average of humans' and agents' predictions. That is somewhat surprising given previous suggestions (e.g. Einhorn, 1972) that mechanical aggregation is

superior to aggregation of predictions done by humans. We suspect that this may be due to the fact that not all participants in our study were experts, and those who were therefore less self-confident gave more weight to agent predictions when they were not sure of their own predictions. Informal interviews with participants also suggest that experts learned how to identify the conditions where agents will be wrong, and thus, knew when to ignore agent input.

Sharpe Ratio Analysis

Results of the Sharpe ratio analysis, done in a similar manner to the analysis of study 1, are summarized in Table 5.

Table 5 – Sharpe Ratio Analysis

Sharpe Ratio

	Scoring Rule		
	AMAE (Benchmark = 0.5)	AMSE (Benchmark = 0.75)	ALSR (Benchmark = 1.699)
Agents	0.31	-0.04	-0.08
Uninformed Humans	0.43	0.26	0.25
Avg. of Humans & Agents	0.44	0.24	0.23
Informed Humans	0.51	0.31	0.27

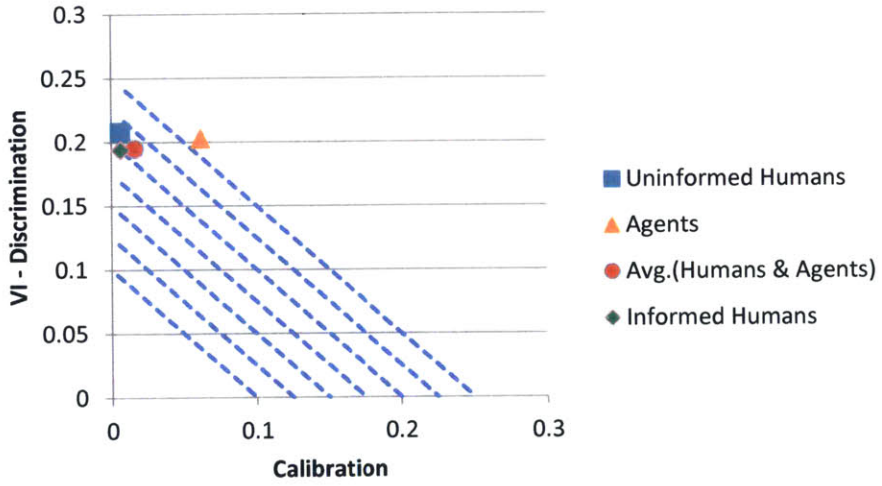
As we can see, under the Sharp index criterion, informed humans outperformed all other conditions, yielding the highest Sharpe ratio under three different scoring rules. Thus, even though the agents' overall performance was worse than that of uninformed humans, humans were

still able to use the agent predictions as valuable data that informed and improved their own predictions. The negative Sharpe ratios for the agents, under the AMSE and ALSR criteria, stem from the number of relatively large errors they have made – yielding a result that is worse than the (completely ignorant) benchmark of always predicting a 50% probability of pass.

Calibration and Discrimination

Calibration results for Study 2 were similar to those for Study 1, so they are not plotted here. In general, the human predictions (both informed and uninformed) were well calibrated, but the agents' predictions were not. In Figure 5 we plot [Variability – Discrimination] vs. Calibration for each predictor, as well as "efficient front" isopleths of MSE. VI in this study was 0.249. It is clear that humans made the predictions more calibrated. Informed humans were only slightly more discriminating than uninformed humans.

Figure 5 – [Variability – Discrimination] vs. Calibration (with MSE isolines. Vi=0.249)



ROC Analysis

The ROC curves of our conditions are presented in Figure 6. Informed humans appear to be more accurate, while agents seem to produce the worst predictions in this study. Areas under the curves are depicted in Table 6.

Figure 6 - ROC Curves for predictions of football plays made by Neural-Net Agents, Uninformed Humans, Average of Uninformed Human and Agent predictions, and by Humans informed of the Agent predictions (55 plays, 10 observations of each play by each condition)

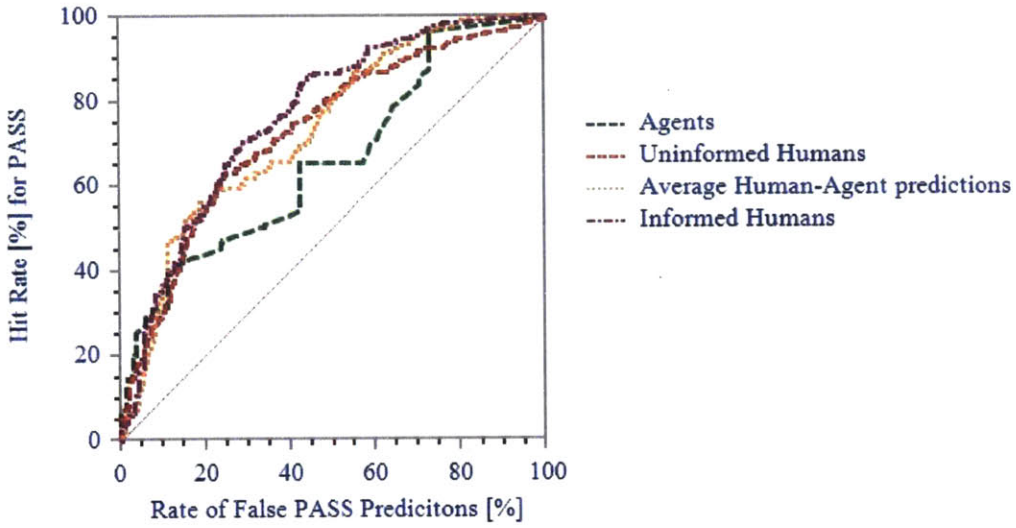


Table 6 - Comparison of the Areas Under the ROC Curves

	Area under ROC Curve	SE[24]
Agents	0.66	0.02
Uninformed Humans	0.73	0.02
Avg. of Humans and Agents	0.73	0.02
Informed Humans	0.77	0.02

This result suggests that the predictions made by humans who were informed of the agent predictions may provide a better trade-off between sensitivity and specificity when compared to predictions made by humans, by agents, or by simply averaging human and agent predictions. In that, it echoes our previous analyses.

Analysis at the Individual Human level

An analysis of the predictions at the individual human participant level reveals a similar pattern to the one we saw at the group level: on average, predictions made by humans after they were informed of the agents predictions were more accurate (MAE: 0.44 before, 0.41 after; MSE: 0.3 before, 0.29 after; LSR: 0.42 before and after). They were more robust as well (Sharpe Ratio based on AMAE was 0.19 before, and 0.25 after; based on AMSE it was -0.15 before and -0.11 after; and based on the ALSR it is -0.23 before and -0.22 after). Areas under the ROC curve are 0.6 before and 0.64 after, with a standard error of 0.005 for either).

On average, as individuals, our participants were not very good predictors. Indeed, the Mean Squared Error and the LSR scores for individual humans are worse than our "ignorant" benchmark. This is also reflected in the respective negative Sharpe-ratio indices. Overall, their predictions show overconfidence, with about a third of the predictions falling in the two most extreme deciles – yet yielding predictions that appear to be both less calibrated and less discriminating than those made by averaging the predictions of several individuals, or by markets of individuals. This is in line with previous work demonstrating the advantage of combining predictions and the superiority of groups to individuals (i.a. Armstrong, 2001; Clemen & Winkler, 1999; Larrick & Soll, 2006; Makridakis, 1989).

Discussion

In two studies, we compared predictions created by combining groups of humans and artificial-neural-net agents to those created by collectives of humans only or agents only. We used 3 different mechanisms to aggregate and combine the predictions: simple-averaging, prediction markets, and letting humans adjust their own predictions based on the agent predictions. We used several different measures and criteria to assess and compare the quality of the predictions, including accuracy (measured using 3 common scoring rules), Sharpe ratios, calibration, discrimination, and receiver-operating characteristic plots.

The combination of humans and agents proved to be overall more accurate than either humans alone or agents alone according to 2 scoring rules that are appropriate for our setting (MSE and LSR). This result holds regardless of the combination mechanism. The combination of humans and agents – in all three forms – provided predictions that were more calibrated than those of the agents, more discriminating than those of the humans, and overall providing a better tradeoff of calibration and discrimination compared to the humans alone or the agents alone. All three forms of combining human and agent predictions also provided the best tradeoff of accuracy and variability of prediction errors, as measured by the Sharpe ratio. In addition, similar results are reflected in the ROC analyses, which do not rely on any assumptions about the cost of errors. In general, therefore, the combination of human and agent predictions in our setting proved to be both more accurate and more robust than either the agents-only predictions or the humans-only predictions.

What would have happened had we used better agents? Likely, we would have gotten more accurate predictions overall. But we expect that

the general scheme of improvement will persist, so long as the incentives are properly aligned. People would tend to intervene less if they note that the markets or the agents are doing a good job of prediction, since they would have less opportunities to earn. While a methodical examination remains for future studies, informal short interviews conducted with several study 2 participants indicate that indeed some of the participants with higher expertise have learned to identify the cases in which agents were right and wrong, and have adjusted their predictions accordingly, while more novice participants tended to rely more on the agent predictions.

What do these results imply about combining predictions of humans and models or agents in general? As previous research has shown, there are many situations where mechanical predictions based on structured data are superior to predictions made by humans (Grove et al., 2000), and in such contexts it may be desirable to use purely automated predictions in these cases. As trends of progress in computing capacity, accumulation of big data, and advancement in artificial intelligence continue, the ability of computers to acquire, process and analyze new types of data (e.g. facial recognition, sentiment-analysis) will grow rapidly, and so will their ability to outperform human experts in some prediction tasks. There are also situations where the underlying dynamics and factors relevant to predicting are so complex, or where there is so little codifiable data, that the only option is to rely on human judgment.

But there are also, still, many important real-world situations where combining predictions from both humans and machines can be valuable. Based on the relative strengths of humans, and machines, we argued that this approach may be particularly beneficial in semi-structured situations, where patterns are difficult to discern, where data are difficult to codify

and/or quantify, or where sudden changes might occur unexpectedly. Our results provide support to this claim.

Our work contributes to the growing body of knowledge about predictions by characterizing a class of prediction problems that can most benefit from combining multiple human and machine predictions, and by empirically demonstrating the advantage of that approach in a representative instance of that hypothesized class. We provide a contemporary example, using machines that run adaptive models based on machine learning, rather than static models used in most previous literature on combining human and model predictions.

We deliberately tested combination methods that are simple to operate, from the point of view of the forecaster, and that are robust to changes in the number and quality of individual forecasters. Previous theoretical and empirical work has shown that no single prediction method, and no single combination mechanism can be universally superior for all cases and circumstances, and that is obviously true for the class of complex prediction problems we discuss here. Rather than making a definitive claim about the superiority of a specific combination mechanism, our work here highlights the potential of improving forecasts by combining predictions from multiple people and machines. That the combined predictions in our setting were best (by multiple criteria), regardless of combination method, provides indication that the approach has merit.

In reality, selection of forecasting and forecasting-combination methods is subject to context-specific considerations that go beyond measures of goodness of predictions. We find prediction markets specifically interesting in the context of combining human and machine

predictions, not only because they provide a seamless and flexible mechanism for incorporating relevant information, specifically broken leg cues, but also, for reasons we mentioned above, that make them appealing in organizational settings when addressing the type of problems we discuss here, such as their usefulness in motivating people with relevant information to share it and participate in the prediction process. Many empirical studies, done in a wide range of settings and contexts, suggest that prediction markets are at least as good as other forecasting, and forecasting-combination methods, and often better, however, we are not aware of previous attempts to use them for combining human and machine predictions. In study 1 we demonstrate that prediction markets can be used effectively in this way.

Some recent work by others offers additional new insights into the art and science of combining predictions. For example, Jain, Mukherjee, Bearden, and Gaba (2013) propose a time-unpacking process, in which unpacking the distal future into intermediate more proximal futures yields wider confidence intervals that help overcome overconfidence, and systematically improves calibration; Lichtendahl, Grushka-Cockayne, and Winkler (2013) show analytically and empirically that averaging quantiles is better than averaging predictions; and in an extensive study of geopolitical predictions, Ungar, Mellers, Satopää, Tetlock, and Baron (2012) achieved improvements by training experts, and by transforming aggregate forecasts, making them more extreme. These were tested on human predictions, but should likely be helpful also when combining predictions from humans and machines, and new studies should attempt to combine these insights with ours.

Conclusion

The explosive growth of Big Data, and recent advances in the field of artificial intelligence and cognitive computing (Kelly & Hamm, 2013) have created a significant shift in how organizations perform tasks, as gradually more tasks previously done by humans are being fully done by machines (Brynjolfsson & McAfee, 2011). As the power of machines to find patterns in data, relate them to the real world and make predictions is dramatically rising, there is room to reconsider the relative roles and work breakdown between humans and computers in some prediction tasks. We believe that for the type of prediction tasks we have identified, human intervention and expertise will remain necessary in the foreseeable future.

Additional work needs to be done to clarify the best ways of making this combination in various specific contexts. Given the rapid pace of change, empirical work is warranted, and we hope our work will encourage others to further investigate this promising direction. Beyond the realm of making predictions, exploring new ways of connecting the knowledge, skill and intelligence residing in people's minds with the power of artificial intelligence may also prove beneficial in other types of tasks performed by individuals, groups and organizations.

Acknowledgments

We are grateful to John Willett for his patience, dedication and help with statistical analyses and to Chris Hibbert for software development, and education about prediction markets. This project originated, and has benefited greatly from discussions with Sandy Pentland, Tomaso Poggio, Drazen Prelec, and Josh Tenenbaum. Jason Carver, Wendy Chang, and Jeremy Lai have developed much of the software used in this project, and together with Rebecca Weiss have greatly helped with designing and running the experiments. For their wise comments, we also thank John Carroll, Gary Condon, Robin Hanson, Haym Hirsh, Ben Landon, Retsef Levi, Cynthia Rudin, and Paulina Varshavskaya. Thanks also go to our undergraduate research assistants – Jonathan Chapman, Catherine Huang, Natasha Nath, Carry Ritter, Kenzan Tanabe and Roger Wong – for their help in running experiments. Partial funding was provided by the MIT Lincoln Laboratory and by the U.S. Army Research Laboratory's Army Research Office (ARO).

References

Abdellaoui, M., Bleichrodt, H., l'Haridon, O., & Paraschiv, C. (2013). Is There One Unifying Concept of Utility? An Experimental Comparison of Utility Under Risk and Utility Over Time. *Management Science, 59*(9), 1-17.

Armstrong, J. S. (1989). Combining forecasts: The end of the beginning or the beginning of the end? *International Journal of Forecasting, 5*, 585-588.

Armstrong, J. S. (2001). Combining forecasts. In J. S. Armstrong (Ed.), *Principles of forecasting: a handbook for researchers and practitioners*: Kluwer Academic Publishers.

Berg, J. E., Forsythe, R., Nelson, F., & Rietz, T. A. (2008). Results from a Dozen Years of Election Futures Markets Research. In C. R. Plott & V. L. Smith (Eds.), *Handbook of Experimental Economic Results* (pp. 742-751). Amsterdam, The Netherlands / Oxford, UK: Elsevier.

Berg, J. E., Nelson, F. D., & Rietz, T. A. (2008). Prediction market accuracy in the long run. *International Journal of Forecasting, 24*(2), 283-298. doi: 10.1016/j.ijforecast.2008.03.007

Blattberg, R. C., & Hoch, S. J. (1990). Database Models and Managerial Intuition: 50% Model+ 50% Manager. *Management Science, 36*(8), 887-899.

Bohorquez, J. C., Gourley, S., Dixon, A. R., Spagat, M., & Johnson, N. F. (2009). Common ecology quantifies human insurgency. *Nature, 462*(7275), 911-914.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78*(1), 1-3.

Brown, R. (1986). *Social psychology* (2nd ed.). New York, NY: Free Press.

Brynjolfsson, E., & McAfee, A. P. (2011). *Race against the machine: How the digital revolution is accelerating innovation, driving productivity, and irreversibly transforming employment and the economy*: Digital Frontier Press Lexington, MA.

Bunn, D. W., & Wright, G. (1991). Interaction of judgemental and statistical forecasting methods: Issues and analysis. *Management Science, 37*(5), 501-518.

Camerer, C. F., & Johnson, E. J. (1991). The process-performance paradox in expert judgment: How can experts know so much and predict so badly? In K. A. Ericsson & J. Smith (Eds.), *Toward a general theory of expertise: Prospects and limits* (pp. 195-217): Cambridge University Press.

Chen, K.-Y., & Plott, C. R. (2002). Information Aggregation Mechanisms: Concept, Design and Implementation for a Sales Forecasting Problem. *California Institute of Technology, Division of the Humanities and Social Sciences, Working Paper 1131.*

Chen, Y., Chu, C.-H., Mullen, T., & Pennock, D., M. . (2005). *Information markets vs. opinion pools: an empirical comparison.* Paper presented at the Proceedings of the 6th ACM conference on Electronic commerce, Vancouver, BC, Canada.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting, 5*(4), 559-583.

Clemen, R. T., & Winkler, R. L. (1999). Combining probability distributions from experts in risk analysis. *Risk Analysis, 19*(2), 187-203.

Clements, M. P., & Hendry, D. F. (1993). On the limitations of comparing mean square forecast errors. *Journal of Forecasting, 12*(8).

Cowgill, B., Wolfers, J., & Zitzewitz, E. (2008). Using Prediction Markets to Track Information Flows: Evidence from Google. *Dartmouth College.*

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*(4899), 1668-1674.

Dawes, R. M., & Kagan, J. (1988). *Rational choice in an uncertain world*: Harcourt Brace Jovanovich San Diego.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics, 44*(3), 837-845.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 253-263.

Duin, R., & Tax, D. (2000, June 21-23). *Experiments with classifier combining rules.* In proceedings of the First International Workshop on Multiple Classifier Systems (MCS 2000), Cagliari, Italy.

Einhorn, H. J. (1972). Expert measurement and mechanical combination. *Organizational Behavior and Human Performance, 7*(1), 86-106.

Ferro, C. A. T. (2007). Comparing Probabilistic Forecasting Systems with the Brier Score. *Weather and Forecasting, 22*(5), 1076-1088.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association, 102*(477), 359-378.

Goel, S., Reeves, D. M., Watts, D. J., & Pennock, D. M. (2010). *Prediction Without Markets*. In proceedings of the 11th ACM conference on Electronic commerce, Cambridge, MA.

Goldberg, L. R. (1970). Man versus model of man: A rationale, plus some evidence, for a method of improving on clinical inferences. *Psychological Bulletin, 73*(6), 422-432.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological), 14*(1), 107-114.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12*(1), 19-30.

Hayek, F. A. (1945). The Use of Knowledge in Society. *The American Economic Review, 35*(4), 519-530.

Hibon, M., & Evgeniou, T. (2005). To combine or not to combine: selecting among forecasts and their combinations. *International Journal of Forecasting, 21*(1), 15-24.

Ho, T. K., Hull, J. J., & Srihari, S. N. (1994). Decision combination in multiple classifier systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 16*(1), 66-75.

Hopman, J. (2007). Using forecasting markets to manage demand risk. *Intel Technology Journal, 11*, 127–136.

Jain, K., Mukherjee, K., Bearden, J. N., & Gaba, A. (2013). Unpacking the Future: A Nudge Toward Wider Subjective Confidence Intervals. *Management Science, 59*(9), 1-18.

Janis, I. L. (1972). *Victims of groupthink: A psychological study of foreign-policy decisions and fiascoes*: Houghton Mifflin Boston.

Janis, I. L., & Mann, L. (1977). *Decision making: A psychological analysis of conflict, choice, and commitment*: The Free Press New York.

Jewson, S. (2004). The problem with the Brier score. *Arxiv preprint physics/0401046*.

Johnson, E. J. (1988). Expertise and decision under uncertainty: Performance and process. In E. J. Johnson (Ed.), *The nature of expertise* (pp. 209-228): Lawrence Erlbaum Associates.

Jolliffe, I. T., & Stephenson, D. B. (2003). *Forecast verification: a practitioner's guide in atmospheric science*: Wiley.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*: Cambridge University Press.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological review, 80*(4), 237-251.

Kelly, J. E., III, & Hamm, S. (2013). *Smart Machines: IBM's Watson and the Era of Cognitive Computing*: Columbia Business School Publishing.

Kittler, J., Hatef, M., Duin, R. P. W., & Matas, J. (1998). On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(3), 226–239.

Kleinmuntz, B. (1990). Why we still use our heads instead of formulas: toward an integrative approach. *Psychological Bulletin, 107*(3), 296-310.

Kuncheva, L. I. (2002). A Theoretical Study on Six Classifier Fusion Strategies. *IEEE Transactions on pattern analysis and machine intelligence, 24*(2), 281-286. doi: 10.1109/34.982906

Lam, L. (2000, June 21-23). *Classifier combinations: implementations and theoretical issues.* In proceedings of the First International Workshop on Multiple Classifier Systems (MCS 2000), Cagliari, Italy.

Larrick, R. P., & Soll, J. B. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science, 52*(1), 111.

Lichtendahl, K. C., Jr., Grushka-Cockayne, Y., & Winkler, R. L. (2013). Is It Better to Average Probabilities or Quantiles? *Management Science, 59*(7), 1-18.

Lichtenstein, S., Baruch, F., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases*: Cambridge University Press.

Littell, R. C., Milliken, G. A., Stroup, W. W., & Wolfinger, R. D. (2006). *SAS for mixed models* (2nd ed.): SAS Publishing.

Makridakis, S. (1989). Why combining works? *International Journal of Forecasting, 5*(4), 601-603.

Makridakis, S., & Winkler, R. L. (1983). Averages of forecasts: Some empirical results. *Management Science*, 987-996.

Malone, T. W. (2004). Bringing the market inside. *Harvard Business Review, 82*(4), 106-114.

Mannes, A., Michael, M., Pate, A., Sliva, A., Subrahmanian, V. S., & Wilkenfeld, J. (2008). Stochastic Opponent Modeling Agents: A Case Study with Hezbollah. In H. Liu, J. J. Salerno & M. J. Young (Eds.), *Social Computing, Behavioral Modeling, and Prediction* (pp. 37-45).

Meehl, P. E. (1954). *Clinical vs. statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis: University of Minnesota Press.

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology, 12*(4), 595–600.

Murphy, A. H., & Winkler, R. L. (1977). Reliability of subjective probability forecasts of precipitation and temperature. *Applied Statistics*, 41-47.

Murphy, A. H., & Winkler, R. L. (1987). A general framework for forecast verification. *Monthly Weather Review, 115*(7), 1330-1338.

Myers, D. G., & Lamm, H. (1976). The group polarization phenomenon. *Psychological Bulletin, 83*(4), 602-627.

Negash, S. (2004). Business intelligence. *The Communications of the Association for Information Systems, 13*(1), 54.

Ortner, G. (1997). *Forecasting Markets—An Industrial Application*. University of Technology Vienna.

Perols, J., Chari, K., & Agrawal, M. (2009). Information market-based decision fusion. *Management Science, 55*(5), 827-842.

Rabin, M. (1996). Psychology and Economics. *Journal of Economic Literature, 36*(1), 11-46.

Schoonjans, F., Zalata, A., Depuydt, C. E., & Comhaire, F. H. (1995). MedCalc: a new computer program for medical statistics. *Computer Methods and Programs in Biomedicine, 48*(3), 257-262.

Seifert, M., & Hadida, A. L. (2013). On the relative importance of linear model and human judge (s) in combined forecasting. *Organizational Behavior and Human Decision Processes, 120*(1), 24-36.

Sharpe, W. F. (1966). Mutual fund performance. *Journal of business, 39*(1), 119-138.

Sharpe, W. F. (1994). The Sharpe ratio. *Journal of portfolio management* (Fall), 49-58.

Spann, M., & Skiera, B. (2009). Sports forecasting: a comparison of the forecast accuracy of prediction markets, betting odds and tipsters. *Journal of Forecasting, 28*(1), 55-72.

Stewart, T. R. (2001). Improving reliability of judgmental forecasts. In J. S. Armstrong (Ed.), *Principles of forecasting: A handbook for researchers and practitioners* (pp. 81-106): Kluwer Academic Publishers.

Suen, C. Y., & Lam, L. (2000, June 21–23). *Multiple classifier combination methodologies for different output levels*. In proceedings of the First International Workshop on Multiple Classifier Systems (MCS 2000), Cagliari, Italy.

Sunstein, C. R. (2005). Group Judgments: Statistical Means, Deliberation, and Information Markets. *New York University Law Review, 80*, 962.

Sunstein, C. R. (2006). *Infotopia: How Many Minds Produce Knowledge*: Oxford University Press, USA.

Surowiecki, J. (2004). *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. New York: Doubleday.

Swets, J. A. (1973). The relative operating characteristic in psychology. *Science, 182*(4116), 990-1000.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240*(4857), 1285-1293.

Swets, J. A., & Pickett, R. M. (1982). *Evaluation of diagnostic systems: methods from signal detection theory*: Academic Press.

Téglás, E., Vul, E., Girotto , V., Gonzalez, M., Tenenbaum, J. B., & Bonatti, L. L. (2011). Pure Reasoning in 12-Month-Old Infants as Probabilistic Inference. *Science, 332*(6033), 1054-1059 doi: DOI: 10.1126/science.1196404

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to Grow a Mind: Statistics, Structure, and Abstraction. *Science, 331*(6022), 1279.

Tetlock, P. (2005). *Expert political judgment: How good is it? How can we know?* : Princeton University Press.

Twardy, C., Hanson, R., Laskey, K., Levitt, T. S., Goldfedder, B., Siegel, A., D'ambrosio, B., & Maxwell, D. (2014). *SciCast: Collective Forecasting of Innovation*. Paper presented at the 2nd Collective-Intelligence Conference, Cambridge, Massachusetts, USA.

Tziralis, G., & Tatsiopoulos, I. (2007). Prediction Markets: An Extended Literature Review. *Journal of Prediction Markets, 1*(1), 75-91.

Ungar, L., Mellers, B., Satopää, V., Tetlock, P., & Baron, J. (2012). *(2012). The Good Judgment Project: A Large Scale Test of Different Methods of Combining Expert Predictions*. In proceedings of the 2012 AAAI Fall Symposium Series.

Winkler, R. L. (1969). Scoring rules and the evaluation of probability assessors. *Journal of the American Statistical Association, 64*(327), 1073-1078.

Winkler, R. L. (1989). Combining forecasts: A philosophical basis and some current issues. *International Journal of Forecasting, 5*(4), 605-609.

Wolfers, J., & Zitzewitz, E. (2004). Prediction markets. *The Journal of Economic Perspectives, 18*(2), 107-126.

Wolfers, J., & Zitzewitz, E. (2006). Interpreting Prediction Market Prices as Probabilities. *CEPR Discussion Paper No. 5676*.

Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry, 39*(4), 561-577.

This page intentionally left blank

Essay 3

Augmenting Collective-Intelligence:

Accelerating the Review of Complex Intellectual Artifacts in Open-Innovation Challenges

Yiftach Nagar

Patrick De Boer

Ana Cristina Bicharra Garcia

Abstract

A critical bottleneck in open-innovation systems is the process of reviewing and selecting the best submissions. This bottleneck is especially problematic in settings where submissions are complex intellectual artifacts whose evaluation requires expertise. To help reduce the review load from experts, we offer a computational approach that relies on analyzing sociolinguistic and other characteristics of submission text, as well as activities of the crowd and the submission authors, and scores the submissions. We developed and tested models based on data from contests done in a large citizen-science platform - the Climate CoLab - and find that they are able to accurately predict expert decisions about the submissions, and can lead to substantial reduction of review labor, and acceleration of the review process.

Introduction

Collective-intelligence systems are increasingly used to elicit intellectual artifacts including ideas, plans, designs and predictions from crowds, in order to address various large-scale challenges. Some notable examples include InnoCentive, which runs ideation competitions on various subjects (Boudreau, Lacetera, & Lakhani, 2011), Galaxy Zoo, in which volunteers help scientists to classify galaxies (Raddick et al., 2010), and FoldIt, in which the crowd assists in predicting protein-folding structures (Cooper et al., 2010). The designers of these systems (and many other) overcame two critical challenges: they have managed to motivate and harness intellectual work of hundreds of thousands of people, using different incentives that appeal to intrinsic and extrinsic motivations. They have also invented novel ways of organizing the work of crowds of people such that inputs from lay people can be validated, refined, and combined with other inputs, to yield outcomes of surprisingly high quality.

However, designers and operators of many collective-intelligence systems face another critical challenge: the *evaluation and selection* of submissions from the crowd. In this paper, we propose a novel computational approach that can help reduce human experts' cognitive load by performing initial screening and/or prioritization of the review queue. Our approach is unique in that it is the first, to our knowledge, to model both the submitted artifact itself and traces of human activity relating to the submission; and in considering how sociolinguistic aspects of the submission text may influence expert reviewers. We tested our approach in context of the Climate CoLab – an open-innovation citizen-science system, and we show it can yield significant improvement in the process.

Focus: crowd innovation challenges

While many problems are routinely solved by individuals or teams within the boundaries of traditional organizations, the knowledge and skill required for solving some very complex and/or novel problems are distributed widely outside the organization, sometimes in places that are not known in advance (Chesbrough, 2003; Malone, Laubacher, & Dellarocas, 2009). Indeed people coming "from the outside", who bring different perspectives and heuristics often generate the best solutions to innovation challenges (Jeppesen & Lakhani, 2010). Turning to Open Innovation and seeking ideas and knowledge widely outside of the organization can help organizations to discover radical new ways to think about their problems, and to solve them in creative ways (Lakhani, Jeppesen, Lohse, & Panetta, 2007; Terwiesch & Xu, 2008).

These realizations led many companies, governments, and non-profits to develop and run crowd-innovation challenges (cf. Boudreau et al., 2011; Boudreau & Lakhani, 2013; Morgan & Wang, 2010; Terwiesch & Xu, 2008). Indeed, in many cases, these efforts yielded solutions to problems where previous attempts have failed (Lakhani et al., 2007). However, receiving a lot of submissions from a large and diverse crowd also created a new challenge: evaluating a mass of complex intellectual artifacts (Nagar, 2013).

The evaluation challenge: the bottleneck of expertise

In some crowdsourcing systems, evaluating crowd inputs is straightforward. For example, reCAPCHA (Von Ahn, Maurer, McMillen, Abraham, & Blum, 2008) uses simple statistical calculations to processes many millions of human OCR entries for obscured scanned words each

day. In FoldIt, domain expertise was coded into the system such that each proposed way of folding a protein can be computationally evaluated in an instant.

However, evaluating intellectual artifacts becomes a significant challenge when those artifacts are complex (e.g. when they contain a lot of unstructured text), when no computational methods are applicable, where domain expertise is scarce, or where criteria for evaluation are not always clear. For instance, the assessment of proposed solutions to scientific challenges posted by InnoCentive requires high levels of expertise that is not readily codifiable, and which is only available in the heads of domain experts; a resource that is both scarce, and expensive. Or, consider the review of NSF grants: according to Boudreau et al. (2012), in 2010 alone the NSF brought over 19,000 scientists to the Washington DC area to participate in proposal evaluation. Beyond the potential to incur significant costs, the bottleneck of expertise for performing evaluation and selection can cause substantial delay in finding solutions. In September 2008, Google launched a crowd-innovation challenge called "Project 10^{100}" that asked people to submit ideas that will change the world. According to Google's original plan, winners were supposed to be announced in January 2009, following a 3-month review cycle. After receiving over 150,000 submissions, Google postponed the announcement of winners multiple times, and eventually, announced them only in September 2010, twenty months(!) after the original planned date. During that time, thousands of Google employees took part in the review process. As these examples clearly illustrate, the bottleneck of expertise for performing evaluation and selection can become one of the most critical hurdles for collective-intelligence systems addressing large problems.

Relieving the Bottleneck of Expertise

Relieving this bottleneck of expertise in the review process is not of the type of challenge that can simply be "solved". But we believe that research and innovation in ways of organizing review work, as well as in complementary computational approaches, such as the one we introduce in this paper, can lead to significant improvement over the current state of the art. Before discussing our approach, we review related work.

Related Work

The current state-of-the-art in reviewing submissions is not much different from the state-of-the-art a half-century ago: in the field, the cumbersome, labor-intensive, slow process of expert panels, review committees, and variations thereof, is still dominant.

More recently, some attempts were made to relieve the bottleneck of expertise for reviewing; but to our knowledge no new method has been widely adopted yet. These attempts generally fall into two categories: organizational (mainly using crowdsourcing), and computational.

Crowdsourcing evaluations

Crowd-based evaluation and filtering systems come in several variations: crowds (which might include organization members, volunteers, or paid crowds in online labor markets) are recruited and are asked to *vote* (e.g. Bao, Sakamoto, & Nickerson, 2011), *rate* (e.g. Blohm, Riedl, Leimeister, & Krcmar, 2011; Riedl, Blohm, Leimeister, & Krcmar, 2013; Salminen & Harmaakorpi, 2012), or *rank* submissions (e.g. Salganik & Levy, 2012), based on one, or multiple criteria (Dean, Hender, Rodgers, & Santanen, 2006). *Prediction markets*, in which crowd members trade

contracts based on their beliefs about the likelihood of ideas to be successful, have also been proposed as a way to incentivize and aggregate ratings from the crowd for predicting the quality or success of ideas (Bothos, Apostolou, & Mentzas, 2012; Soukhoroukova, Spann, & Skiera, 2012).

Although large crowds may be relatively easily recruited to off-load experts, these methods have all been shown (both theoretically and empirically) to have flaws and limitations (see (Klein & Garcia, 2013) for a more nuanced and elaborate review). The main limitation (though not the only one) common to all of these approaches is that the size and complexity of the task of comparing alternatives rises rapidly with the number of options that need to be compared. This renders them problematic for use in practice in crowd-innovation challenges where the number of ideas is large, especially when the proposals are complex.

Computational evaluations

Automatic Essay Scoring algorithms (also known as Auto-graders) have been in use for some time now, and recently received renewed attention with advances in natural language processing, and the growing need for such tools for Massive Open Online Courses (MOOCs) (Markoff, 2013). Indeed, some positive results have been reported (Shermis & Hamner, 2013). However, these tools usually require to be trained on large corpora of manually annotated student essays, which are assumed to be somewhat similar to one another. By the very nature of innovation challenges, submissions are very diverse, and creating annotated sets is a labor intensive, slow process. It is also not clear whether and to what extent the rules developed in some setting will be applicable in other settings (Klein & Garcia, 2013). For these reasons, automatic essay

scoring tools have not been applied to judging submissions in open innovation contests. Another approach, closer to the one we present in this work, is to develop metrics of the quality of submissions based on word frequency statistics. For example, Walter and Back (2013) measured the use of unique sets of words in order to detect innovative ideas, with mixed results. Although we find their approach promising, the average submission in their setting had 25 words only. Therefore, implications from their study for systems in which submissions are more complicated, may be limited. Westerski, Dalamagas, and Iglesias (2013) developed an elaborate domain-independent taxonomy for idea annotation, and offer that idea originality and idea dependability (the level to which it is connected to other ideas), are strong predictors.

These results are encouraging, and yet, we believe that there is room for additional work. We present our method in the following section.

Method

Approach

Our approach was shaped by several realizations regarding different aspects of the problem:

1. Predicting the winners of open-innovation contests is hard and tricky even for experts. An easier approach is to differentiate high quality from low quality submissions. Such a "triage" step is performed manually in many systems, and does not really necessitate high-level expertise to perform. It is probably the place where computational means can achieve the most reliability, and most impact, by filtering out low-quality submissions, and freeing

the experts to devote their time and skill to consider more promising submissions.

2. Although the content of submissions can be computationally modeled in various ways, true understanding of complex ideas is still the realm of humans. However, form matters too, and can be reliably assessed computationally, based on solid theoretical foundations.
3. While expert-panels are notoriously prone to many types of bias, due to a lack of a better alternative their judgment is still de-facto the state of the art, and widely accepted as the gold standard in studies.
4. Previous attempts to computationally classify and rate crowd-proposals relied solely on proposals' text (e.g. Walter & Back, 2013; Westerski et al., 2013). Yet, specifically in open-innovation environments, traces of crowd and author activities in relation to the proposals are available, and may provide additional clues that can help predict which proposals would be favored by expert judges.

Our resulting approach is open-ended and greedy. Based on data available in our setting, we devised a preliminary taxonomy of variables which can serve as a guideline for modeling work, and which can be enhanced and appropriated to fit different settings. We developed and tested models based on this taxonomy, which take into account sociolinguistic and other aspects of proposals' text, as well as author and crowd behavior. With these models we aim to match the reviewers decisions at the first triage stage. We demonstrate our approach in the context of one platform – the *Climate CoLab*.

Setting: the Climate CoLab

The Climate CoLab[25] (Introne, Laubacher, Olson, & Malone, 2011) is a sociotechnical system designed to help thousands of people around the world collectively develop plans for addressing global climate change. The CoLab combines several design elements, including model-based planning and simulation, and a crowdsourcing platform where citizens work with experts and each other to create, analyze, and select detailed proposals for what to do about climate change. As of September 2014, over 200,000 people have visited the Climate CoLab, representing virtually every country in the world, and over 24,000 have registered as members. The main activity under the CoLab is a set of ideation contests on a range of topics, from how to reduce emissions from electric power generation to how cities can adapt to climate change. Past winning proposals have been presented to decision makers in the UN and the US congress, and to potential implementers. In the 2012-2013 set of contests that we analyzed, beyond the announcement of winners in each contest, a grand a prize of USD 10,000 was granted to the proposal that was selected best across all contests.

The current review process at the CoLab

Proposals are submitted on the CoLab's website, using a template which asks authors to indicate the what, where, who and when of the proposal. Proposals can include text, as well as multimedia, and while some proposals are incomplete, or of low quality - many high-level

[25] http://www.climatecolab.org

proposals are submitted as well. To select the best ideas, the CoLab staff has developed an ad-hoc review organization that includes volunteers in two roles: *Fellows* are graduate students and young professionals; and *Expert Judges* – mainly senior faculty, and industry veterans. Fellows and judges are recruited for specific contests, based on their expertise in the contest topic. After the fellows perform initial screening, they, together with the expert judges, select "Semi-finalists". Authors of semi-finalist proposals are given a chance to revise their proposals, and after that, judges and fellows select the finalists, from which winners are later selected. This process is very labor-intensive. For reviewing the set of 2012-2013 contests, about 60 volunteer reviewers (about half of them *fellows*) were recruited.

In addition, the crowd (i.e. the CoLab community of registered users) can make comments on proposals and indicate their support for a proposal by clicking a "thumbs-up" button (akin to the "like" action on online social-networking platforms). During the last phase, the crowd also votes to select the crowd-vote awards.

The dataset

Our dataset is comprised of the entire set of proposals that were submitted to the contests that ran under the Climate CoLab framework in 2012-2013. In total, 369 proposals were submitted in 18 contests, which covered a wide range of both technical and social topics related to dealing with climate change, e.g. the reduction of greenhouse gas emissions from transportation systems, geoengineering to avoid methane feedback and urban adaptation. A complete list of the contests and of the proposals is available on the Climate CoLab website.

Of these proposals, 81 proposals (about 22%) were selected as "Semi Finalists" by the CoLab's fellows and judges. These semi-finalists were reviewed more thoroughly, and their authors received detailed constructive comments, and were given an opportunity to submit revised versions. Following another revision cycle, 59 finalists were selected, and eventually winners were elected from this pool of finalists.

We focused attention on this first stage of the review, i.e. the selection of semi-finalists.

The metrics

These include the data of the proposal itself, and activities of authors, community members and the crowd in relation to the proposal. We grouped those data into six categories as detailed below.

Readability of the proposal text

Numerous studies show that easier reading improves comprehension, retention, reading speed and readers' perseverance (also called depth or persistence: the tendency to keep reading the text) (DuBay, 2007). The ease with which a reader reads a text depends on the reader's skill, knowledge, interest and motivation, as well as on features of the text: its content, style, design and organization (DuBay, 2007). Early in the 20th century, *educators started using vocabulary difficulty and sentence length to predict the difficulty of a text. This has spurred intensive research and development of readability formulae.* These formulae use counts of language variables in a piece of writing in order to provide an index of probable difficulty for readers (Klare, 1974). Many readability formulae have been developed over the years, and while there has been critique about their misuse and

their value in certain applications, DuBay (2007) notes that they *"have proven their worth in over 80 years of research and application"*.

Our first hypothesis therefore is that readability may influence human expert judges as they read proposals, and specifically, that low readability will hinder the proposal's chance of being favored by the judges.

We did not define as a goal to find the "best" readability index that would provide the highest correlation with proposals success. There are literally hundreds of readability indices, but they are better thought of as rough guides than as highly accurate values (DuBay, 2007). To check whether our intuition about readability has merit, we selected four indices that are in very common use in many applications: Flesch-Kincaid Grade Level, Flesch Reading Ease (Kincaid, Fishburne Jr, Rogers, & Chissom, 1975), Automatic Readability Index (ARI) (Kincaid et al., 1975), and the Coleman-Liau Index (Coleman & Liau, 1975).

Writing style: Function words and language style matching

Style words, also known as *function words*, including pronouns (such as *I, you, they*), articles (*a, an, the*), prepositions (*to, of, for*), auxiliary verbs (*is, am, have*), and some other common word categories, account for more than half of the words that occur in human communication (whether written or spoken) (Pennebaker, 2011). While these words convey very little meaning on their own, extensive research demonstrated that by analyzing their use, we can learn about the personality, social skills, honesty and intentions of the people who use them (Pennebaker, 2011).

Further, social psychologists and sociolinguists, who study the use of language in social contexts, have shown that people match their language,

stylistically, to that of other people with whom they are communicating. Researchers have further shown that a reliable index of language style matching can be constructed by using counts of function words. This index also correlates with social-psychological phenomena such as the strength of dyadic relationships, group cohesiveness and group task performance (Gonzales, Hancock, & Pennebaker, 2009; Ireland & Pennebaker, 2010).

It seems plausible that language style might also affect expert reviewers' perception of the proposals, and influence their decisions. While we did not have writing examples from the reviewers that may have allowed us to check the matching between their writing styles and those of proposal authors, the pool of reviewers of the Climate CoLab can be characterized with some common traits: highly educated, highly conscientious, working in academia or in knowledge work. We conjectured therefore, that as a collective, CoLab reviewers tend to have similar stylistic preferences regarding the writing of the proposals. For instance, we hypothesized that they will prefer to see proposals that are written in more 'academic', rather than colloquial style.

It is not easy to directly map such a hypothesis to specific function words a-priori. We took an exploratory approach, and – rather than selecting a subset of words based on any theoretical basis – let the data speak. We used the 2007 version of LIWC (Pennebaker, Booth, & Francis, 2007), which calculates the percentage of words in a text that fall into each of 15 function word categories (several of which overlap hierarchically, e.g., first-person singular pronouns are a subcategory of personal pronouns. See Pennebaker, Chung, Ireland, Gonzales, and Booth (2007) for a complete list of variables and further details).

After an initial run, we eliminated variables that had very low frequencies, to remove any noise created by outliers. We then removed variables whose values in the semi-finalists group were not statistically-significantly different from their values in in the non-semi-finalists set (based on Mann-Whitney's two-tailed test, $\alpha = 0.05$).

We thus narrowed down the list to 15 variables[26], which are depicted in Table 7.

Table 7 Remaining LIWC variables

Category	Variable	Examples
Cognitive Processes	Discrepancy	should, would, could
	Inclusive	And, with, include
Linguistic Processes	Auxiliary verbs	Am, will, have
	Common verbs	Walk, went, see
	Dictionary words	
	Personal pronouns	I, them, her
	Present tense	Is, does, hear
	Total function words	
	Total pronouns	I, them, itself
	Words>6 letters	
Personal concerns	Achievement	Earn, hero, win
	Money	Audit, cash, owe
	Work	Job, majors, xerox
Punctuation	Commas	,
	Dashes	-, –

[26] Although word count is included in LIWC, in our modeling we assigned it in a separate category (proposal length, see below).

Potential indicators of the completeness and maturity of the proposal

We have described above the mixed blessing of asking the crowd to submit ideas. On the one hand, with the right incentives, many more ideas are submitted than would have otherwise, raising the likelihood of finding diamonds in the rough. On the other, since the crowd is diverse, and includes people with different levels of relevant knowledge, skill, and motivation, the quality of submissions varies greatly, and the quality of many submissions may be poor.

We hypothesized that several metrics of the text might signal how much work was put into creating the proposals, and accordingly, the completeness and maturity of the proposal:

1. The number of references (NumReferences): the last section in the CoLab proposal template allows the authors to include references to external sources. We hypothesized that more references can signal that more work was done on the proposal. (The reference lists on CoLab proposals are much shorter than those in academic papers, where such a relation is less likely to hold. While the maximum is 49, the mean number of references in our corpus is 4.11 and the median is 1. About 40% of the proposals had no references in the reference section.
2. The number of hyperlinks (NumHyperlinks)
3. The number of images
4. Whether some sections were left unfilled. The proposal submission interface does not force authors to fill all the sections of the proposal. This is done deliberately, to allow people to submit "half-cooked" ideas, and allow others to respond and assist. As a result,

some proposals remain in this state when judging starts. This would not necessarily disqualify them, and it is possible (though less likely) that a proposal can advance to the semi-finalist stage even if not all its sections are complete. We built a set of dummy variables to indicate whether any section was empty.

Length

Length affects readability, it is related also to style, and can potentially indicate something about proposal maturity. Since length is related to all 3 categories above, we decided to treat it as a separate category. It seems likely that immature proposals, e.g. proposals that have complete sections unfilled, would be shorter. We therefore assumed that very short proposals would have a lower chance of being selected. It is not so clear, however, that the relationship is monotonic. One could assume that very long proposals might be frowned upon, at least by some judges.

Length can be measure by the number of letters, words, sentences and paragraphs. We measured all of them.

Crowd activity

In the early stages of the contest, members of the community can indicate their support for a proposal by clicking a "thumbs-up" button (akin to the "like" action on online social-networking platforms). In addition, members of the community (including fellows and judges), as well as authors, and everyone else, can comment on proposals during all phases of the contest.

We considered the following metrics:

1. Number of comments: one the one hand, it seems likely that more interesting ideas will receive more attention, and drive more engagement, which will be positively correlated with comments. On the other hand, it seems likely that people comment when they see flaws, more so than they do just to say words of support. It was therefore not clear to us whether we will see a strong correlation with the outcome, yet we thought this was worth checking.
2. Number of comments made by experts: we checked whether comments made by members of the review team at an early state before the deadline were correlated with later selection of proposals as semi-finalists.
3. The proportion of comments made by experts.
4. The number of "Likes" the proposal received from the community: The CoLab community is an unusually highly-educated community, and yet, on average, members do not have the same level of expertise as the judges. Although it is likely that expert judges will judge proposals somewhat differently from the average member of the community, we still expect to see some correlation between the "taste" of the community, and that of the judges.
5. Proportion of "Likes": Because our data set includes proposals that were submitted to 18 different contest, we adjusted the number of likes, to control for the differences in the number of proposals across contests, and the number of people interested in them. Some contests drew more proposals and more crowd activity than other contests. The proportion of likes is the number of likes a proposal received, divided by the total number of likes received by all proposals in that contest.

Author actions

1. Number of Days: the number of days left between the initial submission and the submission deadline. We have heard speculation and occasional observations from organizers of the Climate CoLab as well as other prize-bearing crowdsourcing ideation challenges, about strategic behavior of some authors, who submit their proposals close to the deadline, seemingly to prevent others from copying their ideas. One contest organizer conjectured that the best proposals are among the last to be submitted, though he had no supporting data. Although it is indeed the case that most proposals are submitted very close to the deadline, this could be the result of mere procrastination. We therefore decided to check whether there is a relationship.
2. Number of Updates: Once a proposal is submitted, authors can update it (as well as let other members of the crowd to do so) as many times as they want. Can the number of updates help predict which proposal will be selected as a semifinalist? More updates may mean that the proposal was not well thought of in advance, but they could also mean that more work is being done.

Modeling

Our statistical analyses began with non-parametric correlation tests[27] that helped us identify the variables that would be good candidates for our

[27] We used the Kendall-tau correlation test, since not all of the variables are normally distributed in the dataset.

models, as well as to avoid issues of multicollinearity in our models. We then created a series of logistic-regression models for each category of variables, using partial sets of variables that were not strongly correlated with each other, and came up with the best model[28] for each category after eliminating variables that did not have statistically-significant effects on the outcome variable. We then constructed a set of integrated models, which combined the most salient predictors from all categories, and selected the final model. Finally, we validated our model by building a machine-learning classifier and performing a stratified 10-fold cross-validation.

Results

Logistic regression results

In Table 8 we depict a partial summary of our modeling, comparing the final models in each category, and the final combined model. The final model parameters are depicted in more detail in Table 9.

[28] We compared the models based on measures of goodness-of-fit and predictive power, including -2LL, AIC, Pearson Chi-square, and the area under the ROC curve.

Table 8 - Representative logistic regression models from different categories, and final model

	Readability	Writing Style	Maturity	Length	Crowd Activity	Author Activity	Combined
(Intercept)	-2.694***	-0.305	-2.641***	-2.534***	-0.984***	-0.984***	-2.964***
CLI	0.091**						
ppron		-0.229*					-0.379**
verb		-0.081*					
References			0.057**				
Timeline			1.389***				1.284**
NumWords				0.001***			0.0001***
PropLikes					4.129***		3.517**
NumDays						-0.007**	
-2LL	376.91	374.42	356.41	339.39	376.75	376.12	313.04
AIC	380.91	380.42	362.41	343.39	380.75	380.12	323.04
$p>\chi^2$	0.0007	0.0009	1.13E-07	2.55E-12	0.0006	0.0005	1.67E-15
AUC	0.658	0.647	0.733	0.796	0.680	0.637	0.821

Table 9 - Final Model Parameters

						Confidence interval for Odds-Ratio	
Variable	Estimate	S.E.	Wald Z	p	Odds-Ratio	2.50%	97.50%
(Intercept)	-2.964	0.452	-6.551	5.72E-11	0.052	0.020	0.117
ppron	-0.379	0.127	-2.990	0.00279	0.684	0.526	0.865
Timeline	1.284	0.428	2.997	0.00272	3.612	1.634	8.941
NumWords	0.001	0.000	4.881	1.05E-06	1.001	1.001	1.001
PropLikes	3.517	1.301	2.704	0.00686	33.672	2.606	454.905

Accordingly the model equation is:

$$\hat{p}(semifinalist = 1|ppron, Timline, NumWords, PropLikes) = \frac{1}{1 + e^{-(-2.964 - 0.379\, ppron + 1.284\, Timeline + 0.0001\, NumWords + 3.517\, PropLikes)}}$$

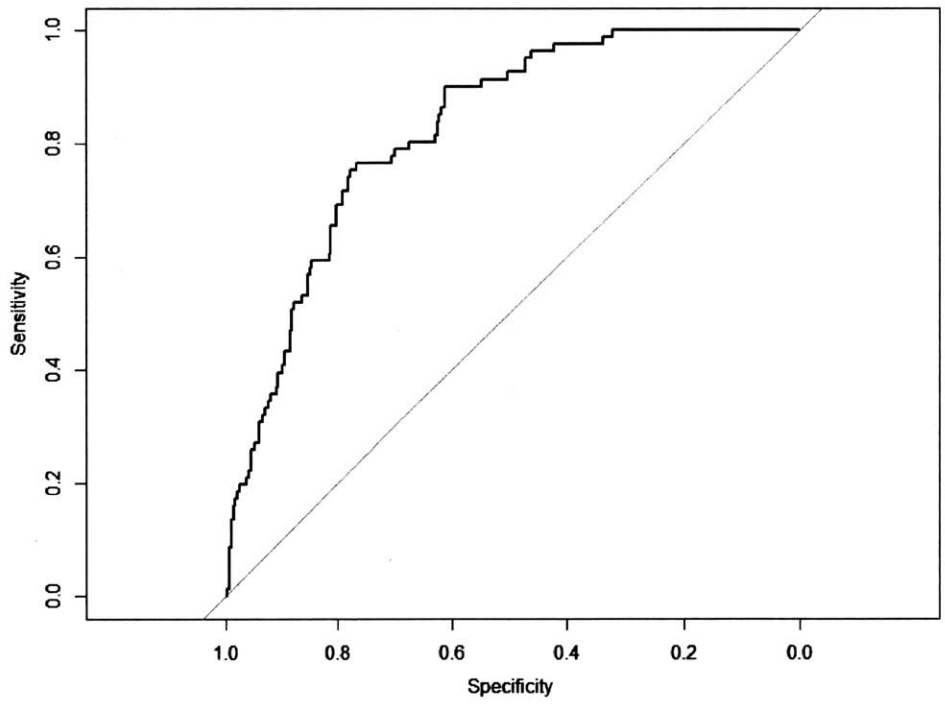
Interpretation of the final model

- **Length** proved to be a very strong predictor: the fitted-odds of a proposal to be selected as a semi-finalist were 2.4 higher if it had 1000 more words.
- **Proposal completeness** is important. Although the CoLab contests yield proposals that experts have deemed as very high quality, many submissions are immature, and even incomplete. The fitted odds-ratio of a proposal which included information in the Timeline section to be selected (compared to a proposal with no timeline) was 3.61. We can also consider the following: of 369 proposals we observed, 112 submissions (about 30%) did not include a timeline (note that the proposal submission template includes a timeline section). Of 81 semifinalists, only 8 were proposals that had no timeline.
- **Pronoun use** was negatively correlated with proposal success: proposals that included more pronouns were less likely to be selected.
- **A higher proportion of "likes"** also indicated higher likelihood for the proposal to be selected, though we should be careful with stating an odds-ratio as this estimate had a large confidence interval. To illustrate, according to the model, a proposal that received 20% of the "likes" in a contest was about 8% more likely to be selected compared to a proposal with no "likes".

Evaluation of model performance, and implications for implementation in practice in the CoLab and in other settings

To validate and evaluate our model, we built a Logistic Regression classifier with scikit-learn (Pedregosa et al., 2011) using the variables we have selected in our combined model, and ran a stratified 10-fold cross-validation (as advised in Kohavi, 1995). The resulting ROC curve is depicted in Figure 7. The resulting model accuracy was 0.789 (±0.06) and the average AUC was 0.816 – very close to the AUC of the combined model in Table 8.

Figure 7 ROC Curve based on stratified cross-validation



We further used bootstrapping to check whether our approach can be used to build a powerful model by using only a subset of previously-judged proposals. The results, in Table 10, are encouraging.

Table 10. Bootstrapping results

% of data used to build the model	Resulting Area under ROC curve
80%	0.81
40%	0.80
5%	0.72

To illustrate the strength, and potential implications of our results, let us consider a couple of examples, by selecting cutoff points on the ROC curve. The selection of such points depends on the preferences of the system designers, i.e. the organizers of each innovation challenge. The nature of many of these challenges is such that the preference of their organizers would be not to miss a good idea that might be "hiding in the haystack". This is often the main motivation for organizing such open challenges in the first place. This is also the case with the Climate CoLab. The cost of reviewing proposals is marginal compared to the benefits we can reap from solutions to problems caused by climate change. That said, operation costs and resources are not unlimited, and the time of expert reviewers, whether they volunteer or they are paid, is a scarce resource, hence the need to reduce the demand for expert time.

Assume, therefore, the case where no good idea should be missed. We will then tune our threshold to maximize sensitivity, on account of specificity. We can select our cutoff rating to be p^=0.033. Results from the model are depicted in Table 11.

Table 11. Model output, threshold = 0.033

	Selected	Not Selected
Indicated by the model	81	236
Not indicated by the model	0	52

The sensitivity of the model in this case is 100% (all semifinalist are indicated by the model). But the model is also able to correctly identify 52 non-finalists. That means, even under the most extreme circumstances, where the model is tuned to 100% sensitivity, it could have reduced the amount of work of experts by filtering out 52 lower-quality submissions (about 14% of all submissions). In a system like the Climate CoLab, which employed over 60 volunteering experts and semi-experts who spent many days reviewing, such a relief can be substantial.

Additional gains can be made by tuning the model to increase specificity. Assume, for example another scenario, where the cutoff point is set on p^=0.185. Results from the model in this case are depicted in Table 12.

Table 12. Model output, threshold = 0.185

	Selected	Not Selected
Indicated by the model	71	112
Not indicted by the model	10	176

In this case, the model sensitivity drops to 87.7%, but specificity goes up substantially: the model would indicate 186 submissions as submissions that are not likely to be selected by expert human judges to be semi-finalists, and would be correct about 176 of those. What about the 10 semi-finalists that we will incorrectly classify as lower quality? When considering the use of a model and cutoff points, we cannot guarantee either 100% sensitivity or specificity. But implementing a prediction

model such as our model can allow contest organizers to organize the review process somewhat differently, and hopefully, more efficiently. For instance, instead of assigning the entire proposal pull to be reviewed by expert judges, they can allocate the review of submissions that were highly-rated by the model (and therefore believed to be more likely to be deemed as higher quality by the judges) to experts, and the review of proposals that were low-rated by the model, to lesser-expert reviewers. In such case, the size of the highly-skilled expert panel can be dramatically reduced (e.g. in the case of the second scenario, the initial pool of submissions sent to high-level experts in this case would have been only 71+112=183. That's a reduction of about 42%, from the 317 proposals we would send to the experts if we choose the more conservative threshold, and of about 50% (!) from the baseline of not using a model at all! Alternatively, contest organizers may keep the review process in the hands of experts, but prioritize the review sequence such that submissions which received lower scores by the model would be reviewed later. This approach can be helpful in cases where the crowd's help is asked in response to crisis, and speed is of essence, such as the crowdsourcing effort the followed the "Deep Horizon" disaster in the gulf of Mexico in 2010.

Discussion

We are proposing a mechanical approach that can assist human-experts by automatically scoring these complex intellectual artifacts. Experts may be able to use these scores as indicators that can assist them in screening the initial pool of submissions, freeing them to dedicate their time to consider the better submissions. Our open-ended approach, which relies on integrating various features of the artifact, with various pieces of

data from human activity relating to the proposal has yielded very promising results when applied to data from the Climate CoLab.

Most of the specific results from our models would not, by and large, be a huge surprise to people who have acquired some experience in running crowd-innovation challenges. They re-affirm some tacit knowledge, e.g. that the quality of a lot of submissions is low; that more complete, more mature proposals, written in a more formal way have a higher chance of being favored by experts. The contribution of this paper lies not in highlighting these relations, but rather in:

1. Suggesting a greedy, open-ended approach that makes use of data from multiple sources, including the artifact and human activity of the crowd;
2. Offering an initial taxonomy that can serve to guide people interested in building predictive models for additional settings, that can lead to significant, tangible improvement in the review process; Specifying many variables in this taxonomy and demonstrating how they can be measured;
3. Empirically demonstrating the implementation of the approach, and its predictive power, using field data from a real live platform.

Additional consideration for practice

The approach we propose should be carefully appropriated, tuned, and tested to fit different settings. Not all variables we have examined may be available or easy to obtain in different settings. Yet, even simplistic and minimal use of some of the measures can be useful. For example, we noted that word count alone would be a very strong predictor in our setting. By

setting a threshold of 100 words, we could eliminate 38 submissions without throwing away any submission that made it to the semifinals. That is already 10% of the submissions! Raising the bar to 250 words would have eliminated 71 proposals (19%), with only one proposals that was later selected as a semi-finalist, and a threshold of 500 words would have indicated 120 proposals (32.5% of the entire pool) as low-quality (with 3 false positives).

One reason to prefer the inclusion of additional variables in the model rather than strictly preferring the most parsimonious model, is that some of these additional variables can help address a concern some contest organizers have regarding fraud-attempts. Because our approach relies on multiple criteria, and since at least some of them are difficult to manipulate (e.g. crowd behavior, and writing style (Ireland & Pennebaker, 2010)) our approach is quite resilient to attempts to game the algorithm. Regarding this concern, it is also worth noting that since we propose to use our method for filtering the lower-quality proposal, and leave the evaluation of good submissions to experts, any attempt to mechanically game the system is also economically senseless, as such submissions would eventually be reviewed by an expert, and dismissed. As our bootstrapping analysis shows, contest organizers do not need to collect data from an entire contest "round" to build a predictive model of submission quality. A small set of rated proposals can be used to build a preliminary model with good performance. Small "batches" of proposals rated highly by the model can then be judged by expert judges, and small batches of proposals which the model rated low, can be inspected by people with lesser-expertise just to make sure no great submissions are dismissed without proper review. Feedback from these human judges can then be used to fine-tune the model, whether by carefully analyzing the

reasons for model errors, or automatically, by implementing a learning scheme into the model (e.g. by building a Bayesian classifier). We intend to further examine this approach in our future work.

Conclusion

The bottleneck of expertise for reviewing a mass of complex ideas submitted to open-innovation platforms is a real and painful problem. As this model of open-innovation, which has already proven useful in finding solutions to hard problems where none existed before, becomes more prevalent, the need to relieve this bottleneck becomes more acute.

Even the most sophisticated artificial intelligence methods available today are still far from being able to reliably review complex artifacts such as some of the submissions in these platforms (or, say, the papers submitted to academic venues). Yet, computational ways can aid human experts in the review process. A complete solution will therefore include computational means; and a better process of dividing the labor between experts, semi-experts, and non-experts. In this paper we demonstrated an open and greedy computational approach, which leverages multiple available data from the submission text, as well as from traces of human activity relating to the submission. We borrowed different analytical frameworks, mainly from sociolinguistics, and demonstrated how they can be used to guide the development of predictive models for the task at hand. Importantly, while our models proved to be powerful in the real-life setting of the Climate CoLab, focusing on the specific variables that ended up being salient in our models will miss the message of this paper. Our goal here is to suggest an approach, and report encouraging empirical results. We intend to continue working on modeling proposal success in the Climate CoLab, and to refine this model by observing additional years,

and by taking additional variables into account. Further modeling work in different settings will help strengthen the external validity of our results, and provide insight regarding which variables, or families of variables, are important to look at in any settings, and which are context specific. We hope our work will encourage others to join us in pushing open innovation forward.

Acknowledgments

We are grateful to the our mentors and colleagues at the MIT Center for Collective Intelligence and at The Climate CoLab – Tom Malone, Gary Olson, Jeff Nickerson, Rob Laubacher, Laur Fisher, and Mark Klein – and to Abraham Bernstein and James Pennebaker, for their invaluable support and advice. We also thank Klemens Mang, who provided great research assistance.

References

Bao, J., Sakamoto, Y., & Nickerson, J. (2011). *Evaluating Design Solutions Using Crowds*. In proceedings of the Seventeenth Americas Conference on Information Systems, August 4th-7th.

Blohm, I., Riedl, C., Leimeister, J. M., & Krcmar, H. (2011). *Idea evaluation mechanisms for collective intelligence in open innovation communities: Do traders outperform raters*. In proceedings of the Proceedings of 32nd International Conference on Information Systems.

Bothos, E., Apostolou, D., & Mentzas, G. (2012). Collective intelligence with web-based information aggregation markets: The role of market facilitation in idea management. *Expert Systems with Applications, 39*(1), 1333-1345. doi: http://dx.doi.org/10.1016/j.eswa.2011.08.014

Boudreau, K. J., Guinan, E., Lakhani, K. R., & Riedl, C. (2012, December 4). *The Novelty Paradox & Bias for Normal Science: Evidence from Randomized Medical Grant Proposal Evaluations*. Harvard Business School Technology & Operations Mgt. Unit Working Papers, (Working Paper No. 13-054). SSRN(Available online at SSRN: http://ssrn.com/abstract=2184791).

Boudreau, K. J., Lacetera, N., & Lakhani, K. R. (2011). Incentives and problem uncertainty in innovation contests: An empirical analysis. *Management Science, 57*(5), 843.

Boudreau, K. J., & Lakhani, K. R. (2013). Using the Crowd as an Innovation Partner. *Harvard Business Review, 91*(4), 60-69.

Chesbrough, H. W. (2003). Open Innovation: The New Imperative for Creating and Profiting from Technology (Hardcover) *Harvard Business School Press Books*.

Coleman, M., & Liau, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology, 60*(2), 283.

Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., Leaver-Fay, A., Baker, D., Popovi, Z., & Foldit_Players. (2010). Predicting protein structures with a multiplayer online game. *Nature, 466*(7307), 756–760.

Dean, D. L., Hender, J. M., Rodgers, T. L., & Santanen, E. L. (2006). Identifying quality, novel, and creative Ideas: Constructs and scales for idea evaluation. *Journal of the Association for Information Systems, 7*(10), 646-698.

DuBay, W. H. (2007). *Smart Language: Readers, Readability, and the Grading of Text*: ERIC.

Gonzales, A. L., Hancock, J. T., & Pennebaker, J. W. (2009). Language style matching as a predictor of social dynamics in small groups. *Communication Research*.

Introne, J., Laubacher, R. J., Olson, G. M., & Malone, T. W. (2011). *The Climate CoLab: Large scale model-based collaborative planning*. In proceedings of the Conference on Collaboration Technologies and Systems (CST 2011), Philadelphia, PA.

Ireland, M. E., & Pennebaker, J. W. (2010). Language style matching in writing: Synchrony in essays, correspondence, and poetry. *Journal of personality and social psychology, 99*(3), 549-571. doi: 10.1037/a0020386

Jeppesen, L. B., & Lakhani, K. R. (2010). Marginality and problem-solving effectiveness in broadcast search. *Organization Science, 21*(5), 1016-1033.

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. *DTIC Document*.

Klare, G. R. (1974). Assessing Readability. *Reading Research Quarterly, 10*(1), 62-102. doi: 10.2307/747086

Klein, M., & Garcia, A. C. B. (2013). *High-Speed Idea Filtering With the Bag of Lemons*. SSRN. (Available online at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2331563).

Kohavi, R. (1995). *A study of cross-validation and bootstrap for accuracy estimation and model selection*. In proceedings of the International Joint Conference on Artificial Intelligence.

Lakhani, K. R., Jeppesen, L. B., Lohse, P. A., & Panetta, J. A. (2007). *The Value of Openness in Scientific Problem Solving*. Harvard Business School Working Papers, (07-050). (Available online at

Malone, T. W., Laubacher, R. J., & Dellarocas, C. (2009). *Harnessing Crowds: Mapping the Genome of Collective Intelligence*. MIT Sloan Working Papers, (CCI Working Paper 2009-001). (Available online at http://hdl.handle.net/1721.1/66259).

Markoff, J. (2013, April 5). New Test for Computers: Grading Essays at College Level. *The New York Times*.

Morgan, J., & Wang, R. (2010). Tournaments for Ideas. *California management review, 52*(2), 77-97.

Nagar, Y. (2013, Feb 24). *Designing a Collective-Intelligence System for Evaluating Complex, Crowd-Generated Intellectual Artifacts*. In proceedings of the 2013 ACM Conference on Computer Supported Collaborative Work (CSCW 2013), San-Antonio, Texas, USA.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825-2830.

Pennebaker, J. W. (2011). The secret life of pronouns: How our words reflect who we are: New York, NY: Bloomsbury Press.

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. (2007). The development and psychometric properties of LIWC2007. Austin, TX: LIWC. Net.

Raddick, M. J., Bracey, G., Gay, P. L., Lintott, C. J., Murray, P., Schawinski, K., Szalay, A. S., & Vandenberg, J. (2010). Galaxy Zoo: exploring the motivations of citizen science volunteers. *Astronomy Education Review, 9*(1).

Riedl, C., Blohm, I., Leimeister, J. M., & Krcmar, H. (2013). The Effect of Rating Scales on Decision Quality and User Attitudes in Online Innovation Communities. *International Journal of Electronic Commerce, 17*(3), 7-36.

Salganik, M. J., & Levy, K. E. C. (2012). *Wiki Surveys: Open and quantifiable social data collection*. (arXiv:1202.0500v1). Arxiv(Available online at http://arxiv.org/abs/1202.0500v1).

Salminen, J., & Harmaakorpi, V. (2012). Collective Intelligence and Practice-Based Innovation: An Idea Evaluation Method Based on Collective Intelligence *Practice-Based Innovation: Insights, Applications and Policy Implications* (pp. 213-232): Springer.

Shermis, M. D., & Hamner, B. (2013). Contrasting State-of-the-Art Automated Scoring of Essays. In M. D. Shermis & J. C. Burstein (Eds.), *Handbook of Automated Essay Evaluation: Current Applications and New Directions* (pp. 313): Routledge.

Soukhoroukova, A., Spann, M., & Skiera, B. (2012). Sourcing, filtering, and evaluating new product ideas: an empirical exploration of the performance of idea markets. *Journal of Product Innovation Management, 29*(1), 100-112.

Terwiesch, C., & Xu, Y. (2008). Innovation Contests, Open Innovation, and Multiagent Problem Solving. *Management Science, 54*(9), 1529-1543. doi: doi:10.1287/mnsc.1080.0884

Von Ahn, L., Maurer, B., McMillen, C., Abraham, D., & Blum, M. (2008). reCAPTCHA: Human-based character recognition via web security measures. *Science, 321*(5895), 1465.

Walter, T. P., & Back, A. (2013). *A Text Mining Approach to Evaluate Submissions to Crowdsourcing Contests.* In proceedings of the 46th Hawaii International Conference onSystem Sciences (HICSS).

Westerski, A., Dalamagas, T., & Iglesias, C. A. (2013). Classifying and comparing community innovation in Idea Management Systems. *Decision Support Systems, 54*(3), 1316-1326. doi: http://dx.doi.org/10.1016/j.dss.2012.12.004