Resilient Institutions and Social Norms: Some Notes on Ongoing Theoretical and Empirical Research

Forthcoming in Rice, M. (ed.) *Merrill Series on The Research Mission of Public Universities*

John Symons
Professor of Philosophy,
Director of the Center for Cyber-social Dynamics
University of Kansas

Stacy Elmer
PhD Candidate
Department of Philosophy
University of Kansas

"Community resilience" describes the capacity to withstand and bounce back from an adverse event or perturbation. The term is most often used in reference to the ability of a community to recover from disruptions caused by terrorist attacks or natural disasters. One of us (Stacy Elmer) was partly responsible for disaster recovery in President Obama's White House, and among the responsibilities of this role was the development of national-level policies on resilience, in the context of cybersecurity, natural disasters, and bioterrorism preparedness.[1] Inevitably, our societies are subject to a variety of significant threats, and it is prudent to assume that we will simply be unable to prevent all disruptions. Thus, cultivating and supporting resilience has become a high priority for responsible leaders. Government, industry, and charitable organizations have increasingly focused programming and funding aimed at community resilience. However, as we learn more about the kinds of disruptions and threats faced by the United States, it becomes clear that the concept of resilience itself needs to be carefully rethought.[2]

In this paper we review some of the reasons for refocusing on social determinants of resilience rather than on physical infrastructure. Much of the resilience of our societies is due to cultural and normative factors that have generally escaped attention in research on resilience. Most obvious perhaps is the role of social institutions in community resilience. We also argue that traditional approaches to the ontology of critical social institutions miss the role of social norms in the constitution and maintenance of institutions. The resilience of institutions, we argue, is

[1] Stacy Elmer served as Director for Incident Management in The Obama White House, National Security Staff where she managed the National Exercise Program portfolio, including the development and coordination of senior-level disaster response exercises. She led the Interagency Policy Sub-Committee on Exercises and Evaluation, contributed to the development of national-level policies on resilience, cybersecurity, and bioterrorism preparedness and provided incident management for the President during disaster response.

[2] While studies show that there is no evidence of a common definition of community resilience, nine core elements that are common to the idea of resilience have been identified: local knowledge, community networks and relationships, communication, health, governance and leadership, resources, economic investment, preparedness, and mental outlook.

dependent on associated social norms. Once we see the role of social norms in institutions, we can recognize that those norms pose a potential vulnerability that can become an attack surface for adversaries. The paper closes by considering some of the ways that our adversaries can undermine adherence to social norms and some of the ways that such attacks might be studied empirically.

What is resilience?

Let's begin with the basics: Given the notorious vagueness of the concept of resilience, an initial reaction might be to say that it is not the kind of concept that really has any significant empirical content. While many of the standard definitions are subject to criticism, we can offer a rough list of characteristics that capture the important features associated with resilience without attempting a philosophically rigorous definition in terms of necessary and sufficient conditions.

A system can be said to be resilient if it:

- is prepared for intervention or perturbation.
- maintains its identity and bounces back after attack.
- adapts in ways that are guided by its identity in a timeframe that is appropriate to its identity.
- learns from past perturbation or intervention.

To say that some community is resilient or that one community proved more resilient than another has some significance insofar as it seems to mark something about the properties of cities, communities, or institutions that is responsive to empirical reality. We seem to correctly recognize that some features of social reality have the capacity to endure in ways that others don't. In other words, in spite of metaphysical or ontological scruples, our capacity to rank some systems as more resilient than others in virtue of evidence from both ordinary experience and scientific inquiry is sufficient to ground further investigation.

It's also the case that we can misjudge the resilience of some social systems. The fact that we can be surprised by or proven wrong about the resilience of a system counts as some evidence that resilience should be understood realistically. It's often remarked that the resilience of the Soviet system in the 1980s was overestimated and that the resilience of the global financial system was underestimated in the period following the 2008 financial crisis. Our judgments of the resilience of social systems can be corrected by the course of history, but our intuitive sense that there is something or some cluster of things that makes some systems resilient and others fragile seems reasonably clear and certainly sufficient to warrant further investigation.

While much of our research focusses on highly theoretical aspects of the philosophy of social science, we recognize that policy making operates at a different timescale than philosophy. Policy makers do not have the luxury of waiting until fundamental questions in social ontology are settled.[3] Instead, they are tasked with making practical and often urgent decisions concerning

[3] Elsewhere, one of us has argued in detail as to why scientific inquiry into emergent properties like the resilience of social systems need not wait for the metaphysical status of non-fundamental properties to be established (Symons, 2018).

the resilience of critical social systems. To date, on our view, policy making around resilient communities and institutions has largely focused in the wrong place. Our theoretical work aims to correct this error by refocusing debates around community resilience on genuinely social aspects of communities and institutions rather than on either physical infrastructure or individual psychology.

From Physical Infrastructure to Social Infrastructure

Historically, discussions of resilience have tended to focus on the underlying physical infrastructure (roads, power grids, water sources, etc.) supporting the basic functioning of a community. Network measures and features have been a primary means of measuring resilience in, for example, computer and telecommunications networks (Modarresi and Symons, 2021; 2020a; 2020b). While there is a vital role for these approaches, especially in the study of engineered systems, we have argued that the bottom-up approach to resilience is inadequate (Pipa and Symons, 2019). Bridges, roads, and power grids are built and maintained by complex social institutions.[4] If those social institutions fail to perform their role, physical infrastructure quickly disintegrates. Thus, there is a top-down role for social infrastructure in relation to physical infrastructure. In fact, an intact and resilient society can generally rebuild physical infrastructure or successfully adapt to its loss. By contrast, a society in which critical institutions have failed and is thereby unable to solve problems collectively will be unable to maintain complex physical infrastructure.

Over the past 10 years it has become clear that the focus on the physical basis of resilience must change (Patel et al., 2017). On the one hand, researchers increasingly recognize the importance of cultural factors and social relationships in the resilience of communities. On the other hand, the use of social media in malicious interventions by adversaries of the United States has forced attention to the vulnerability of social institutions and social norms. This new attention has widened our understanding of the factors affecting the resilience of communities. Cyberattacks on institutions involved in banking, healthcare, and commerce have likewise drawn attention to the role of the non-physical, social relations and institutions that play a critical role in community resilience.[5] Attacks on Google (2009), RSA (2011), JP Morgan, (2014), the Ukrainian power grid (2015), etc., all leveraged social engineering hacks either via phishing/spear-phishing emails, by telephone (voice phishing), or by gaining physical access through the use of a deceptive pretext or via physical media. Understandably, given the prominence of these hacks, security science has focused attention on the vulnerability of individuals. However, our work aims to encourage a new focus on the distinctively social aspects of the social attack surface, rather than on interventions targeting individual beliefs or attitudes.

For the remainder of the paper, we will sketch some of the core issues around community resilience in the context of traditional theories concerning the nature of institutions. These debates are interdisciplinary in nature, involving history, political science, sociology, economics, and

[4] Thanks to Bert Westbrook for pressing us on this point.

[5] Even in cases where our focus involves modeling physical infrastructure around, for example, food, energy and water, social and economic factors are increasingly recognized as directly relevant. See, for example, Modarresi & Symons (2021).

anthropology, and within philosophy these debates cut across subdisciplines such as social ontology, political philosophy, and game theory. While we cannot survey all aspects of the debate, we argue that one of the most prominent contemporary views of institutions in economics and philosophy, the rules and equilibria approach, fails to account for the resilience of social institutions. Currently, there is no good explanation of why and how some institutions are more or less resilient. We argue that examining institutional failure provides a useful way to understand what makes a social system resilient, and offers a way to explain how resilience can be cultivated within our communities.

Understanding what kinds of things count as real and what criteria we use to decide such questions falls to a branch of philosophy known as ontology.[6] Our work assumes a stance towards basic questions in social ontology that we do not defend in detail in this paper. However, our ontological commitments are guided by a commonsense attitude towards our policy responsibilities. Practical decisions concerning, for example, defensive measures in the social attack surface would be impossible if we took the view, for example, that there is "no such thing as society." The denial of the role of the social in interstate conflict would be a grave mistake even if we do not have a well-grounded theory of the ontology of social phenomena ready to hand.

The study of the ontology of social things, for example, money, nations, communities, institutions, etc., is the domain of a subdiscipline known as social ontology. Social ontology explains the structure of social reality by exploring how social entities exist and relate to other things in the world. Institutions comprise a key focus of social ontology. Social ontology studies the function that institutions play in society and the reasons for their existence. Our view is that explanations of resilience in social systems raise questions about the nature of institutions to which social ontology must respond. Our approach is broadly consonant with the position defended by Brian Epstein (2015). We argue in particular that social ontology, especially when it attempts to tackle social properties like resilience should not take an individualist methodological strategy.

Rules vs. Equilibria

Early investigations into the nature of the firm in the 1960s by Ronald Coase helped to frame later debates concerning the ontology of institutions. Coase suggested that firms function to lower transaction costs that would otherwise be incurred in forming contracts among individuals (1990, p. 3-13). He noted the role of institutions like firms in economic processes, but other economists sought to give an account of the nature of the firm itself rather than its role in the broader economic system. For example, Douglass North argued for a rules-based conception of institutions in which rules serve to structure political, economic, and social interactions in society; institutions are the codification of these rules that shape human behavior in the "game of society" (1991, p. 97-112). According to North, institutions function to improve the welfare of society by enabling human beings to achieve their goals. The trouble with North's account for our purposes is that it does not explain why some rules are followed while others are not. While

[6] In analytic philosophy, ontology has taken a variety of forms in the 20th and 21st centuries. For an overview see Symons (2010).

rules are clearly constitutive of institutions in some important respects, some account of the relationship between human decision-making and rules is required.

Game theoretic approaches to institutions attempt to respond to concerns of this kind. Most prominently, David Lewis' equilibria account of institutions applies game theoretic principles to explain why human beings follow the rules that comprise institutions. Lewis describes institutions as behavioral patterns that human beings settle into within a society and suggests these patterns can be explained as solutions to coordinated games with multiple equilibria (2008, p. 21). Actions that are in equilibrium will be repeated in the course of many actions because they are stable, while choices that reflect non-cooperative equilibria are unstable and thus unlikely to be repeated (Lewis, 2008, p. 42).

We share a version of Avner Grief's criticism of Lewis' account. If institutions are simply equilibria in a coordination game as Lewis suggests, then rules would not be necessary for establishing institutions (Grief, 2006, p. 12). Grief points out that rules play a vital role in shaping behavior so as to reach the equilibria that form institutions by acting as strategies that *ought* to be followed. Rules are statements of what ought to be done; they specify what behaviors are expected, which in turn creates regularities in behavior that people recognize and use to condition their own behavior (Grief, 2006, p. 15). This "rules-in-equilibrium" approach recognizes that there are incentives for people to follow the rules that are established by an institution and is an attempt to ground the relationship between human agency and institutional structures from an individualistic perspective (Grief, 2006, p. 211).

What is missing in these analyses is the role of social norms in relation to institutional rules. In practice, social norms determine whether people in fact follow institutional rules. For example, laws against corruption in bureaucratic life exist in most countries. Whether such rules are followed is a matter of the social norms that operate in those societies and these vary widely. John Searle's "constitutive rules" account is another prominent approach to the ontology of institutions that in some respects is closer to our view. However, as we shall see, Searle's account also misses the essential role of social norms in institutions. For Searle, institutions are systems of constitutive rules. Constitutive rules are those that take the form "X counts as Y in C" where X is a brute fact, Y is an institutional fact, and C is the context in which the institutional fact is accepted (Searle and Willis, 1995, p. 44). Brute facts are those for which no explanation is possible.[7] Institutional facts are those that exist only in the context of human institutions. Institutional facts exist only because human beings believe them to exist, communicate them, and act in accordance to these beliefs. For this reason, language is critical to institutional facts. According to Searle, institutions exist only because people believe them to exist.

On Searle's account, in order for institutional facts to exist there must be a system of constitutive rules that govern their existence. A constitutive rule differs from a regulative rule (one that follows the form "do X" or "if Y, do X") in that regulative rules regulate preexisting forms of behavior or activities that exist independent of the rule (e.g., imperatives); these activities

[7] See Symons (2019) for a discussion of the relationship between brute facts, scientific explanation, and ontology.

or behaviors are logically independent of the rules. Constitutive rules constitute the activities or behaviors that they regulate; these activities or behaviors logically depend on the rules. Thus, constitutive rules constitute new forms of activities or behaviors; they create institutional facts (Searle, 2018, p. 51-54). A constitutive rule for money might go something like "these pieces of paper" (X) count as "money" (Y) in the United States (C).

Institutions can be understood as systems of constitutive rules, or the rules that a person must follow (or follow at least a large subset of) to be considered to be participating in the activity. In the case of money, the constitutive rules are the rules that comprise the recognized system for exchange within a society. For an institution to exist this system of constitutive rules must continually be recognized and accepted by a sufficient number of people within a society. This recognition creates what Searle calls a "status function." A status function is the power that human beings collectively attribute to certain things. Searle thinks of these as "deontic powers" such as rights, duties, obligations, requirements, and entitlements (2010, p. 224).

For an institution to exist it must have a status function, and for a status function to exist there must be a status function declaration, which is a verbal declaration that communicates a social practice that is recognized and accepted by persons within a society. Assigning a status function to a brute fact signifies the acceptance of that institution (Searle 1995, p. 34). Searle further differentiates between kinds of rules. A status rule defines the meaning of a status given to a thing, while a base rule spells out the conditions a thing must have to achieve that status. The status rule for money is that money is a means of exchange, while the base rule is money must be a piece of paper printed in a specific way by a specific entity.

According to Searle a status function can be represented as a constitutive rule (a rule of the form "X counts as Y in C"). The "counts as" component of the formula is where the status function does the critical work, as the function cannot be achieved by the brute fact (X) alone. In addition, for an institutional fact to exist, collective recognition and acceptance is critical only for the function associated with the brute fact. For the pieces of paper to count as money people have to recognize and accept the function: these pieces of paper function as a means of exchange. It is not necessary for people to recognize that a function has been attributed to a physical substance (e.g., these pieces of paper had no economic value until they were assigned as valuable means of exchange) only that these pieces of paper (money) are a means of exchange.

According to Searle once a status function acquires collective acceptance and becomes a general policy it gains a normative status and becomes a constitutive rule (1995, p. 48). It is normative because people acknowledge that there are behaviors that align (thus also behaviors that do not align) with the rule; in other words, there are established ways to both follow and not follow the rule. Searle's reliance on individual belief in his social ontology of institutions has encouraged a focus on an individualist and epistemically focused understanding of influence campaigns against critical social infrastructure. Our perspective emphasizes the role of social rather than individual epistemic factors in norm adherence. It is common for people with exotic beliefs to act in ways that are in adherence to social norms and are not disruptive to the institutions with which they interact. By contrast, when norms erode, even agents with Searle-style beliefs about the institution and its rules will act in ways that undermine the institution. Thus, we can see the limitations of an account of institutional resilience that relies on epistemic states of individual agents. One's beliefs concerning the existence of a constitutive rule are distinguishable from the likelihood that one will adhere to the rule. This is where social norms play a central role in bringing institutions to life.

To this point, variations on two main approaches to the ontology of institutions have been explored: the rules-based approach and the equilibrium approach. The rules-based approach positions institutions as behavioral rules that guide and constrain behavior during social interaction, while the equilibrium approach treats institutions as equilibria of strategic games. Francesco Guala's theory of institutions falls somewhere between Grief's rules-in-equilibrium approach and Searle's constitutive rules approach and is designed to show that Searle's constitutive rules approach can be encompassed within the rules-in-equilibria approach (Hendriks and Guala, 2015).

Guala accepts that institutions guide, and in some circumstances mandate, people's behavior, which he believes also aligns with our intuitive understanding of institutions (2016). However, he contends that the rules approach does not provide an account of why some rules are followed while others are not. From the equilibrium approach he endorses the idea that successful institutions are comprised of rules that people are motivated or are incentivized to follow (Guala, 2016, p. 10). Incentives can be represented by strategic games, specifically coordinated games with multiple equilibria. In these games, each equilibrium represents a solution to a problem of coordination where the beliefs and behaviors of people are mutually consistent. This latter point is important – not all equilibria are institutions. If an equilibrium can be reached without any player correlating their strategy with the strategy of any other player, then it fails to be an institution.

Institutions require human interaction, and as such require correlation devices. Not all real-world circumstances mirror coordination games with symmetric equilibria. In games with asymmetric equilibria where one of the players must accept to a lower payoff, there has to be some way of coordinating the actions of the players. A correlation device serves this purpose by acting as a signaling mechanism. A traffic light (green means go, red means stop) is an example of a correlation device. Although all players may wish to pass through the intersection first, they recognize that the best move for everyone is to abide by whichever light they happen to arrive to (red or green). In this way correlation devices lead to correlated equilibria.

However, not all correlated equilibria are institutions. For example, non-human animals use correlation devices to signal certain behaviors in certain circumstances. A male seal protects his harem by barking loudly in the water. If another male seal approaches the rookery and hears this barking it will retreat. If it doesn't hear this barking it will proceed. This is an example of non-human animals using a correlation device to solve a coordination game. However, this correlated equilibrium requires that the seal use one strategy, coordinated through a specific signal that dictates a specific behavior. The stimulus (sound of barking) is coupled with the behavior (retreating). For human beings the social world is filled with a multitude of signals and correlation devices that can be decoupled by creating representations.

Representations enable people to draw on a multitude of equilibrium strategies in symbolic form to determine the best course of action and to create new equilibria. Rules are simply symbolic representations of the strategies that ought to be followed in a game (Guala, 2015). Rules serve to coordinate behavior by stipulating behavioral patterns that can be expected of everyone. Rules represent equilibria (in some cases multiple rules together represent correlated equilibria where each rule is a strategy and the equilibrium are the set of strategies/rules). While the rules are general and accepted by all players, each rule/strategy will be followed by a particular player depending on the specific circumstances in which they find themselves.

To this point we have explored the idea that institutions function to provide solutions to coordination games and drive actions of people towards these solutions through institutional rules. Institutional rules create the rights and obligations that dictate how people should or must act in specific circumstances, and in this respect have deontic powers. Unlike Searle, Guala does not think a joint commitment to follow the rules is required for effective institutions because the main role of institutions is to drive solutions to coordination problems. Thus, all people need is concordant expectations about one another's behavior, which are built from both public signals and social interaction.

Guala also modifies Searle's account of a constitutive rule. Recall that according to Searle constitutive rules comprise institutions by governing human behaviors in societies. Constitutive rules are normative when they are collectively recognized, and they correspond to rights and obligations that dictate actions that people can/must perform in certain situations. Such rules are effective only if there are incentives that motivate people to follow them.

Searle's account of constitutive rules takes the form:

X counts as Y in C

Where X represents a brute fact, Y represents an institutional fact, and C is the context in which the institutional rule is accepted.

Guala revises this statement in the form:

If C then X is Y, and if Y then Z

He does this by translating "X counts as Y" to "X is collectively accepted as Y" and interpreting "is collectively accepted as" to "is," resulting in the translation of "counts as" to "is" (Guala, 2016):

counts as ↔ is collectively accepted as
counts as ↔ is collectively accepted as ↔ is
counts as ↔ is

In the money example, certain pieces of paper count as money, thus certain pieces of paper are collectively accepted as money, thus certain pieces of paper are money.

certain pieces of paper count as money ↔
certain pieces of paper are collectively accepted as money ↔
certain pieces of paper are money

This approach undermines Searle's concept of a status function. Recall that a status function is assigned when there is collective acceptance of the purpose of a certain thing (such as money having the status function of being a means of exchange). Guala eliminates the role of the status function by distinguishing between a *status rule* and a *base rule*. Status rules focus on defining what it means to possess that status (e.g., if the status is money, the status rule is money is a means of exchange). They are the rules that define the behaviors that come with that status, including the rights and obligations. A base rule defines the conditions of acceptance, or what is needed, to possess that status (e.g., pieces of paper or discs of metal printed by the U.S. mint); they are concerned with the ontological basis of the status (Hindricks and Guala, 2015). Thus the base rule is "certain pieces of paper are money in the United States"; which applies today in the U.S. because money in the United States is certain pieces of paper that collectively are accepted as a means of exchange.

Guala takes the "counts as" component of Searle's constitutive rule and relates it to what is needed to possess the status (base rules). If "X counts as Y" and "counts as is equivalent to the conditions of acceptance, then X are the conditions of acceptance for Y, where Y is the content of the status function (aka a status rule). Then these two pieces comprise the following constitutive rule:

If C then X is Y, and if Y then Z

Where "if Y then Z" is a status rules that enumerates the actions that are made available to people. Searle claims that this process of transforming constitutive rules into regulative rules enables the introduction of institutional terms, such as money, property, or marriage, which (when they have collective acceptance) contain a wealth of information about the presuppositions for the conditions of the terms. These terms provide an efficient explanation of the sets of strategies that presuppose institutions. In this sense institutions are symbolic representations of equilibria that are denoted by the term (e.g., money, property, marriage) used to describe the institution. Thus, constitutive rules are linguistic transformations of regulative rules, which rely on a new term being introduced that is used to name the institution (Hindricks and Guala, 2015, p. 473). Hindricks and Guala claim that this transformation shows that the rules-in-equilibrium approach and the constitutive rules approach are consistent.

Furthermore, Guala claims that this unified account aligns the concepts of multiple realizability and multiple equilibria. Multiple realizability, or the idea that multiple iterations of the same property, in this case base rules, can occur in different contexts (e.g., pesos, dollars, and gold nuggets are all collectively accepted as money in different contexts) or that in one context there may be a base rule that describes characteristics that satisfies more than one X-term (e.g., coins and pieces of paper are both money in the United States), which is consistent with multiple equilibria in game theory. Thus institutions are defined by the types of strategic problems they solve, and the types of strategic problems are identified by their function (e.g., institution of money: gold nuggets are money because they fulfill some of the classic functions of money). While Guala's theory of institutions provides useful explanatory power for understanding the functional role of institutions, it does not provide an account robust enough to explain or predict why some institutions are resilient in some contexts but fail in others.

Wlodek Rabinowicz objects to Guala's general rules-in-equilibrium account on the grounds that it (1) excludes morality and other non-instrumental forms of action that do not seem to be in equilibrium and (2) does not account for critical components of institutions, such as the physical properties that comprise them (Rabinowicz, 2018). Rabinowicz distinguishes rules that one is motivated to follow from rules that one *ought* to follow, noting that the former is generally less stringent than the latter, which comprise the requirements of morality. Since Guala claims that institutions are systems of rules in equilibrium, Rabinowicz notes that systems of moral rules are not always in equilibrium and therefore systems of moral rules do not constitute institutions. For Guala, morality is not a particular kind of institution. Moral rules are normative elements of institutions. Since individuals' decisions to adhere or not adhere to norms often results in rewards or punishments, on this view moral rules motivate behaviors by signaling how individuals *ought* to act. In this way norms make human actions more predictable and promote cooperation in circumstances where behaviors would otherwise have been motivated by self-interest. Since institutional rules facilitate coordination in situations where human behavior is unpredictable, and norms make behaviors more predictable, Guala infers that norms are institutional rules that facilitate coordination.

As described, norms are not limited to a particular set of contexts, but instead present in all institutions. Framed as changes in the way incentives are structured, norms do not pose problems for the rules-in-equilibrium framework as Rabinowicz suggests. Instead, norms shift the equilibria of games. A set of actions that is in equilibrium of a game with only self-interested payoffs may be out of equilibrium when norms are considered as a part of the rules; actions that result in self-interested payoffs are not always considered moral ways of acting.

Rabinowicz also objects to Guala's theory on the grounds that Guala defines institutions too narrowly by limiting their scope to the systems of rules that govern them. He argues that Guala makes a *pars pro toto* mistake by taking one aspect of institutions (rules) as representative of the whole, leaving out the material components (buildings, people, etc.) that also comprise systems of institutions. Guala responds that physical properties are still a part of institutions but are secondary to the rules. The rules are the elements of an institution that are essential to comprising its function. Physical materials may exist without a system of rules, but without rules material objects do not serve the functions that comprise institutions. For example, the institution of money sets the rules for when a person can exchange certain pieces of paper for goods. While the people, pieces of paper, and goods exchanged are necessary components to the functioning of the institution of money, these physical elements are of secondary importance to the rules that determine how people are able to exchange these pieces of paper for goods. Money functions as a means of exchange. There must be some material object (pieces of paper, round pieces of metal, gold nuggets, etc.) to participate in the exchange but the specific object is irrelevant (Guala, 2018).

According to Guala, the physical objects, such as pieces of paper in the case of money, serve as correlation devices helping coordinate the actions of the people making the exchange. When understood this way, the rules-in-equilibrium approach acknowledges material components as necessary but not sufficient for the establishment of institutions. Whether the physical object for the institution of money is a piece of paper or a gold nugget, the rules are of primary importance because they define the function (a means of exchange) of the object (piece of paper or gold nugget). Guala's theory of institutions rests on the idea that the primary work of social ontology is to understand the functioning of institutions in general, not to explicate the ontology of institutional objects. It also requires allowing that abstract game theoretic models can capture the functional essence of a particular institution by accepting the idea that there can be a definite set of activities that comprise an institution, such as money or marriage (token/type distinction).

While Guala responds to Rabinowicz criticism from morality by reference to self-interest and payoffs in equilibrium games, this strategy misses the role of non-moral norms. Social norms around corruption, for example, can be distinguished from the moral beliefs that people in corrupt societies might have about corruption. As Bichierri notes, social norms around corruption will generally trump the moral views of their participants. I might know that it is morally wrong to bribe the official, but I also expect that everyone does it and that no one would criticize me too harshly for doing it.

Another way to understand the role of norms is to think about the kinds of things that would bring down or destroy an institution and work backwards from there. Take the institution of academic grading—the institution of grading is assigning marks that reflect the quality of students' work. If faculty were paid different amounts based on the grades they assigned (e.g., $1,000 for every A, $10 for every C) the marks would no longer signal the academic value of the student's work. Instead grading would signal wealth, rather than academic excellence, and would destroy the institution of grading.

The fact that grading is a non-mercenary or a non-market service or transaction is a *constitutive feature* of the institution of grading. This feature was not part of the rules that established the institution, not because it couldn't be written into the rules of grading, but because it is effectively unnecessary to write it in.

Some constitutive features of institutions can be distinguished from the rules that are written to establish those institutions. Knowing only the rules that constitute an institution is not enough to know what that institution is; there are norms that are not written into the rules that must also be understood. In the case of grading, what grading is depends on certain kinds of norms being in place that cannot be found in the rules (e.g., *grading is non-mercenary* was not included in the rules when the institution was established). The rules that establish an institution are a different kind of thing than the function or the norms that constitute the institution.

It is also true that a single violation of these norms does not destroy an institution. If one faculty member or even a group of faculty members take bribes for grades, the institution of grading will not be destroyed. However, if enough faculty violate the norm and grading now signals wealth instead of the quality of a student's work, then the institution of grading is destroyed.

As another example, consider the concept of friendship. You cannot pay for friendship because doing so would undermine the conditions for friendship. Paying someone to be your friend does not actually make them your friend, and the monetary transaction undermines the institution of friendship. This does not mean that friendship does not have value or that you could not put a price on friendship, as you can sacrifice other goods for the sake of friendship, but the relationship of friendship itself is not constituted via market transactions. When examining the positive rules or norms that characterize the maintenance of friendship, the notion that friends cannot be purchased need not figure explicitly. In some sense it goes without saying. However, this constitutive feature of friendship reveals itself upon examination of the things that could destroy the relationship.

Both of these examples illustrate our perspective that understanding the ways that an institution can be destroyed provides meaningful insight into the foundation of that institution beyond what the consideration of rules or equilibria alone can offer. If we want to build resilience into institutions and/or systems of institutions, then we must think about more than just the rules and equilibria. In thinking about institutions and non-market values, for example, it clearly makes no sense to reduce institutions to optimization games or some collective emergent calculation of coordinated interest. Our ongoing research aims to understand how interventions at the level of social norms can undermine institutions. Our assumption is the focusing on ways that institutions can fail will helps us to understand how they are constituted and what makes them resilient.

Interventions Aimed at Disrupting Social Infrastructure

In the foregoing discussion of theoretical work on the nature of institutions we have emphasized the constitutive role of norms. If we have correctly characterized the role of norms then we can begin to ask a set of empirical questions concerning the resilience of institutions. For example, what would count as an attack on the social infrastructure of the United States? Consider the ongoing use of social media platforms by the intelligence agencies of the Russian

Federation. These platforms are widely recognized to have allowed low-cost, deniable, distributed, highly networked, and asymmetric interventions on the social infrastructure of the United States (NATO, 2020). While there are effective methods of tracking the means by which disinformation and propaganda are cultivated by the Russian defense establishment, we do not fully understand whether and how Russia intervenes against critical social infrastructure.[8] If it is the case that our adversaries target social institutions, evaluating and measuring the effectiveness of those interventions is a significant challenge.

At present, the nature and efficacy of different attacks are typically understood in individualist and epistemic terms focusing on measures of political dysfunction such as affective polarization and increased instances of contentious politics. This approach has value, but lacks the broader, system-level analysis of how social relations and norms are harmed and how those harms affect critical social institutions. Typically, indicators for the effects of social attacks are measures of either polarization or the growth of polarized online communities that are imputed to be the result of social media influence campaigns. Given our view of the role of social norms in social institutions as discussed above, we regard traditional focus on disinformation and misinformation as an excessively narrow approach to measuring Russian interventions on the social attack surface. Research into the efficacy of defensive strategies to counteract attacks on social infrastructure is in its early stages (Courchesne, Inglehart, & Shapiro, 2021).[9] Our ongoing work focuses specifically on social norms in order to sketch strategic and practical capacities to understand and defend against social attacks.[10]

The extent to which Russian authorities are intentionally targeting the social and cultural resilience of their adversaries is obviously unknown. Nevertheless, figures from the Russian military establishment have explicitly and publicly connected cultural considerations to their cyberwarfare efforts for over two decades.[11] Moreover, Russian philosophers and intellectuals, most notoriously Alexander Dugin, have regularly framed international relations in terms of competing cultural and spiritual values with varying degrees of strength and resilience. The extent to which such expressions can be understood as indicating strategic military principles is highly debatable. Nevertheless, in December 1996 Chief of the Russian General Staff General Viktor Nikolaevich Samsonov publicly observed that:

The high effectiveness of information warfare systems in combination with highly accurate weapons and nonmilitary means of influence makes it possible to disorganize the system of state administration, hit strategic installations, and affect the mentality and moral spirit of the

[8] The U.S. Government has a dedicated center for countering foreign disinformation, the Global Engagement Center (GEC) at the U.S. Department of State. In a 2020 report entitled *Pillars of Russia's Disinformation and Propaganda Ecosystem*, the GEC outlined the major components of Russian disinformation campaigns. This document provides an excellent overview of the official, proxy, and unattributed communication channels that Russia uses to create and amplify false narratives.

[9] The Carnegie Endowment's Partnership for Countering Influence Operations provides analysis of studies done thus far and has identified significant gaps in understanding and prescriptive measures to combat influence efforts.

[10] The Carnegie Endowment has usefully gathered much of the existing research here: https://carnegieendowment.org/specialprojects/counteringinfluenceoperations#latestAnalysis.

[11] The theory and practice of Russia's diverse approach to communication technology for information warfare and influence operations is well documented (see RAND, 2022).

population. In other words, the effect of using these means is comparable with the damage resulting from the effects of weapons of mass destruction (Grovsdev, 2012).

These comments indicate that the Russian military establishment has at least considered the cyberwarfare role of normative and cultural interventions.[12] Russia's efforts to cultivate grievances and amplify the forces of contentious politics illustrates Russia's use of cyber-enabled information operations as another domain, alongside air, land and sea, to attack adversaries. Contemporary studies have relied on individual-level theories and conceptual frameworks to understand these attacks. For example, Edwards et al. (2017) represent the mainstream view that the "social engineering attack surface is the totality of an individual or a staff's vulnerability to trickery. Social engineering attacks usually take advantage of human psychology: the desire for something free, the susceptibility to distraction, or the desire to be liked or to be helpful." Our approach focuses on the social, rather than the individual. Instead of inferring social consequences from psychological operations at scale, we analyze efforts to undermine norms critical to social infrastructure (Mckay and Tenove, 2021). To this end, we aim to test the hypothesis that Russian attacks aim broadly at the likelihood of adherence to two particular kinds of expectations in a relevant population. These two kinds of expectation are theorized by Christina Bicchieri to undergird adherence to social norms (2016). These are *empirical expectations*: The prediction that people typically act in accordance with the norm, and *normative expectations*: The prediction that people in the relevant community typically judge a norm violator to be blameworthy in some way. Our interdisciplinary approach aims to uncover the specific social mechanisms targeted within such operations.

Our ongoing research begins with a bounded case to chart Russian efforts to influence the "defund the police" discussions from 2019 to the present, in order to determine whether those efforts functioned as interventions in the social norms of the United States. These norms might include, for example, respect for and trust in law enforcement, norms around cooperation with police, reliance on police, and expectations with respect to interactions with police officers. We use a combination of data drawn from Twitter and content from newspapers to explore the dynamics of these interventions. Newspaper content data allows us to document the changing nature of public discourse concerning policing. Twitter data allows us to identify both sources of Russian influence and document how and whether empirical and normative expectations are influenced, evidenced through content propagation, engagement data (sharing, etc.), and the formation of online communities around expressed positions on norms.

While Russia's efforts in social media interventions have been mapped and described by the Global Engagement Center at the Department of State (GEC, 2020), a comprehensive analysis of specifically normative interventions is still ongoing. The reason that we target normative investigations is because we assume (as argued above on theoretical grounds) that social norms

[12] The *Doctrine of Information Security of the Russian Federation* emphasizes "applying information technologies for the preservation of cultural, historical, spiritual and moral values of the multi-ethnic people of the Russian Federation" (Russian Ministry of Foreign Affairs, 2016) and "neutralizing the information impact intended to erode Russia's traditional moral and spiritual values" (2016).

are at least partly constitutive of institutions and that institutions can be undermined by destabilizing social norms. While research has identified disinformation as an increasingly *participatory* and social act, emerging from social networks exposed to influence efforts, it is less clear how (and whether) adversaries act to undermine social norms.[13] This set of challenges has been identified by the Department of Homeland Security as a whole-of-society issue and we believe that our efforts to measure and analyze the social aspects of these interventions is a step towards addressing this.[14]

Twitter now permits access to its historical and real-time data archive.[15] Together with collaborators April Edwards, Deborah Pfaff, and Craig Hayden, we use text mining applications on the Twitter archive in relation to known influence campaigns on social media. The data we hope to generate will allow us to test our hypothesis concerning targeting efforts directed towards normative expectations as described above. Among the strategies that we use are text mining of key phrases involving social knowledge, i.e., "everybody knows," "no one thinks," "[some social group] knows…" and related terms.[16] The computational text analysis approach will be designed to identify and capture "social and cultural concepts."[17] When found together with relevant key words, hashtags, and known Internet Research Agency accounts, we count these as instances of a social norm intervention. The diffusion of these interventions (and the IRA-driven amplifications via retweets, bots, etc.) can be tracked through time, and the main lines of transmission beyond Russian-controlled accounts can be observed.

In addition to tracing the dynamics of these interventions, we use newspaper content to examine the extent to which observed shifts in normative expectations are evidenced in subsequent media framing within national U.S. news coverage. Our research has focused initially on norms around trust in police in the United States from early 2020 to the present. We make use of the existing corpus of Black Lives Matter related tweets (Giorgi et. al, 2021) in addition to tweets related to "Defund the Police" discussions. The initial goal will be to determine whether Russian efforts are explained in terms of the theoretical framework we have described. This work is ongoing and we hope to be able to report back to future Merrill Seminars.

Conclusions

The United States is an open, diverse, and liberal society and, as a result, has a more limited range of defensive options available for the defense of our social institutions as compared with our autocratic adversaries. At present, U.S. laws prevent social media companies from being held liable for content posted on their platforms. Our research will inform options for both

[13] For example, see the study of Russian "participatory propaganda," funded by the Office of Naval Research, Kate Starbird, Ahmer Arif, and Tom Wilson, "Disinformation as Collaborative work: surfacing the participatory nature of strategic information operations." https://dl.acm.org/doi/pdf/10.1145/3359229

[14] https://www.dhs.gov/sites/default/files/publications/ia/ia_combatting-targeted-disinformation-campaigns.pdf

[15] https://developer.twitter.com/en/use-cases/do-research

[16] For background on the logic of social or collective aspects of epistemic phenomena, see Rendsvig and Symons (2021).

[17] This methodology differs from traditional sentiment analysis approaches and is necessary, given the research objective. See Dong Nguyen, Maria Liakata, Simon DeDeo, Jacob Eisenstein, David Mimno, Rebekah Tromble, Jane Winters "How We Do Things With Words: Analyzing Text as Social and Cultural Data" *Front. Artif. Intell.*, 25 August 2020 Sec. Language and Computation. https://doi.org/10.3389/frai.2020.00062

practical defensive measures and regulations in response to interventions that are targeted to harm critical social institutions. Nevertheless, we must learn the full scope by which our adversaries threaten our political and social order in order to develop countermeasures that are effective and comport with our values. At this point, the extent to which adversaries manage to successfully target social norms is unknown.

In principle, as we have shown, social infrastructure is as important to national security as physical infrastructure, and national defense requires that we understand the norms, expectations, and choice architectures (especially at the cyber-social interface) that constitute social institutions. Defense of our nation no longer depends just upon *national* security, but also *human* security—which includes the weakening of social norms and, subsequently, institutions by our adversaries. On a theoretical level, this work contributes to our understanding of the relationship between social norms and institutions. This is a topic of great interest in economics, sociology, and political science. We are also hopeful that work of this kind can help to move Security Studies away from an excessively individualist focus in the study of the social attack surface towards recognition of the role of social norms in interstate rivalry.

Acknowledgments
We would like to thank audiences at the Merrill Research Seminar in Kansas City, as well as valuable feedback and critical commentary from April Edwards, Craig Hanks, Amir Modarresi, Deb Pfarr, and Bert Westbrook. Thanks also to Mike Denning, Mary Hill, and Adam Taggert for support in the development of these ideas. This work has been partially supported by funding from The National Security Agency through the Science of Security initiative contract #H98230-18-D-0009 and the National Science Foundation grant 1856084 for the FEWtures project (PI: Mary Hill).

References

Aceves, W. J. (2018). Virtual hatred: How Russia tried to start a race war in the United States. *Mich. J. Race & L.*, *24*, 177.

Arif, A., Stewart L., & Starbird, K. (2018). Acting the Part: Examining Information Operations Within #BlackLivesMatter Discourse, 2 *Proceedings of the ACM on Human Computer Interaction*, *20*.

Amadae, S. (2011). Normativity and Instrumentalism in David Lewis' *Convention*. *History of European Ideas*, *37*, no. 3, 325-35.

Aydinonat, N., & Ylikoski P. (2018). Three Conceptions of a Theory of Institutions. *Philosophy of the Social Sciences*, *48*, no. 6, 550-68.

Bedau, M., & Humphreys, P. (2008). *Emergence: Contemporary Readings in Philosophy and Science.* MIT Press.

Bicchieri, C. (2005). *The Grammar of Society: The Nature and Dynamics of Social Norm*s. Cambridge University Press.

Bicchieri, C. (2016). *Norms in the wild: How to diagnose, measure, and change social norms*. Oxford University Press.

Binmore, K. (2015). Institutions, Rules and Equilibria: A Commentary. *Journal of Institutional Economics, 11*, no. 3, 493-96.

Boyd, R. (1999). Kinds as the "Workmanship of Men": Realism, Constructivism, and Natural Kinds. *Rationalität, Realismus, Revision, 52-89*.

Bradshaw, S., DiResta, R., & Miller, C. (2022). Playing Both Sides: Russian State-Backed Media Coverage of the #BlackLivesMatter Movement. *The International Journal of Press/Politics*. doi:10.1177/19401612221082052

Brennan, G., & Philip P. (2004). *The Economy of Esteem: An Essay on Civil and Political Society*. OUP Oxford.

Coase, R. H. (1995). The Nature of the Firm. In *Essential Readings in Economics, 37-54*. Springer.

Courchesne, L., Ilhardt, J., & Shapiro, J. (2021, September 13). "Review of Social Science Research on the Impact of Countermeasures Against Influence Operations." *Harvard Kennedy School Misinformation Review*. https://doi.org/10.37016/mr-2020-79

Day, Jamison M. (2014). "Fostering Emergent Resilience: The Complex Adaptive Supply Network of Disaster Relief." *International Journal of Production Research, 52*, no. 7:1970-88.

Edwards, M., Larson, R., Green, B., Rashid, A., & Baron, A. (2017). Panning for gold: Automatically analysing online social engineering attack surfaces. *Computers & Security*, *69*, 18-34.

Epstein, B. (2015). *The Ant Trap: Rebuilding the Foundations of the Social Sciences*. Oxford University Press.

Epstein, B. (2016). "Replies to Guala and Gallotti." *Journal of Social Ontology, 2*, no. 1:159-72.

Farrell, H., & Schneier, B. (2018). Common-knowledge attacks on democracy. *Berkman Klein Center Research Publication*.

Gauss, G. (2010). *The Order of Public Reason: A Theory of Freedom and Morality in a Diverse and Bounded World*. Cambridge University Press.

Gerber, T. P., & Zavisca, J. (2016). Does Russian propaganda work? *The Washington Quarterly*, *39*(2), 79-98.

Giles, K. (2017). Countering Russian Information Operations in the Age of Social Media. *Council on Foreign Relations*.

Giorgi, S., Guntuku, S. C., Himelein-Wachowiak, M., Kwarteng, A., Hwang, S., Rahman, M., & Curtis, B. (2022). Twitter Corpus of the #BlackLivesMatter Movement and Counter Protests: 2013 to 2021. In *Proceedings of the International AAAI Conference on Web and Social Media,* Vol. 16, pp. 1228-1235.

Greif, A. (2006). *Institutions and the Path to the Modern Economy: Lessons from Medieval Trade*. Cambridge University Press.

Guala, F. (2010). "Reciprocity: Weak or Strong? What Punishment Experiments Do (and Do Not) Demonstrate." University of Milan Department of Economics, Business and Statistics Working Paper, no. 2010-23.

Guala, F. (2016). *Understanding Institutions: The Science and Philosophy of Living Together*. Princeton University Press.

Guala, F. (2018). Replies to Critics. *Philosophy of the Social Sciences, 48*, no. 6, 630-45.

Guala, F., and Hindriks, F. (2015). A Unified Social Ontology. *The Philosophical Quarterly, 65*, no. 259:177-201.

Gvosdev, N. K. (2012). The Bear Goes Digital: Russia and Its Cyber Capabilities. *Trans. Array Cyberspace and National Security: Threats, Opportunities, and Power in a Virtual World. Derek S. Reveron.* Washington DC: Georgetown University Press.

Hauswald, R. (2018). "Institution Types and Institution Tokens: An Unproblematic Distinction?" *Philosophy of the Social Sciences, 48*, no. 6:594-607.

Heath, J. (2008). Following the Rules: Practical Reasoning and Deontic Constraint. OUP USA.

Helmus, T. C., Bodine-Baron, E., Radin, A., Magnuson, M., Mendelsohn, J., Marcellino, W., & Winkelman, Z. (2018). *Russian social media influence: Understanding Russian propaganda in Eastern Europe*. Rand Corporation.

Hindriks, F., & Guala, F. (2015). "Institutions, Rules, and Equilibria: A Unified Theory." *Journal of Institutional Economics, 11*, no. 3, 459-80.

Hodgson, G. M. (2018). “Understanding and Defining Institutions: The Contribution of Francesco Guala.” Taylor & Francis.

Hendrix, J. (2018, October 24). Two New Reports Expose How Black Americans are Targeted by Russian Disinformation, *Just Security*). https://www.justsecurity.org/

Lee, D. (2018, February 16). The Tactics of a Russian Troll Farm. BBC. https://www.bbc.com/news/technology-43093390

Lewis, D. (2008). *Convention: A Philosophical Study*. John Wiley & Sons.

Linvill, D. L., Boatwright, B. C., Grant, W. J., & Warren, P. L. (2019). “THE RUSSIANS ARE HACKING MY BRAIN!” Investigating Russia’s internet research agency twitter tactics during the 2016 United States presidential campaign. *Computers in Human Behavior*, *99*, 292-300.

Mäkelä, P., Hakli, R., & Amadae, S. (2018). “Understanding Institutions without Collective Acceptance.” *Philosophy of the Social Sciences*, *48*, no. 6:608-29.

McKay & Tenove. (2021). “Disinformation as a threat to deliberative democracy.” *Political research Quarterly*, *74*, 3:703-717.

Modarresi, A., & Symons, J. (2021). Modeling Resilience for Sustainable Food, Energy, and Water Systems. In *AGU Fall Meeting Abstracts,* Vol. 2021, pp. IN45F-0501.

Modarresi, A., & Symons, J. (2020). Resilience and technological diversity in smart homes. *Journal of Ambient Intelligence and Humanized Computing*, *11*(12), 5825-5843.

Modarresi, A., & Symons, J. (2020). Technological heterogeneity and path diversity in smart home resilience: A simulation approach. *Procedia Computer Science*, *170*, 177-186.

NATO Science & Tech Trends 2020-2040, Exploring the S&T Edge (p. 32 – the Inforsphere).

North, D. C. (1991). *American Economic Association*. The Journal of Economic Perspectives, 5, no. 1: 97-112.

Pacherie, E. (2011). “Framing Joint Action.” *Review of Philosophy and Psychology*, *2*, no. 2:173-92.

Patel, S. S., Rogers, M. B., Amlôt, R., & Rubin, G. J. (2017). “What Do We Mean By ‘Community Resilience’? A Systematic Literature Review of How It Is Defined in the Literature.” PLoS currents 9.

Phillips, W. (2015). *This is why we can't have nice things: Mapping the relationship between online trolling and mainstream culture*. MIT Press.

Pihelgas, M. (2015). Mitigating risks arising from false-flag and no-flag cyber attacks. *CCD COE, NATO, Tallinn*.

Pipa, F., & Symons, J. (2019). Towards an understanding of resilience with complex networks. In *Proceedings of the 6th annual symposium on hot topics in the science of security* (pp. 1-1).

Rabinowicz, W. (2018). “Are Institutions Rules in Equilibrium? Comments on Guala’s Understanding Institutions.” *Philosophy of the Social Sciences, 48*, no. 6:569-84.

Rendsvig, R. & Symons, J. (2021). “Epistemic Logic.” *The Stanford Encyclopedia of Philosophy* , Edward N. Zalta (ed.). https://plato.stanford.edu/archives/sum2021/entries/logic-epistemic/

Rozenshtein, A. (2019) *No, Facebook and Google Are Not State Actors.* https://www.lawfareblog.com/no-facebook-and-google-are-not-state-actors

Searle, J. (2010). *Making the Social World: The Structure of Human Civilization*. Oxford University Press.

Searle, J. (2008). “Language and Social Ontology.” *Theory and Society, 37*, no. 5:443-59.

Searle, J. (2015). “Status Functions and Institutional Facts: Reply to Hindriks and Guala.” *Journal of Institutional Economics, 11*, no. 3:507-14.

Searle, J. (2018). “Constitutive Rules.” *Argumenta*, *4*, (1):51-54.

Searle, J. and Willis, S. (1995). *The Construction of Social Reality*. Simon and Schuster.

Starbird, K., Arif, A., & Wilson, T. (2019, November). “Disinformation as Collaborative work: surfacing the participatory nature of strategic information operations.” *Proceedings of the ACM on Human-Computer Interaction*, Volume 3, Issue CSCW, Article 127. pp 1–26. https://doi.org/10.1145/3359229

Stewart, L. G., Arif, A., Nied, A. C., Spiro, E. S., & Starbird, K. (2017). Drawing the lines of contention: Networked frame contests within #BlackLivesMatter discourse. *Proceedings of the ACM on Human-Computer Interaction*, *1*(CSCW), 1-23.

Sugden, R. (2004). *The Economics of Rights, Co-Operation and Welfare*. Springer.

Symons, J. (2019). Brute Facts About Emergence. In Vintiadis, E., and Mekios, C. eds. *Brute facts*. Oxford University Press.

Symons, J. (2018). Metaphysical and scientific accounts of emergence: Varieties of fundamentality and theoretical completeness. *Emergent Behavior in Complex Systems Engineering: A Modeling and Simulation Approach, 2-20*.

Symons, J. (2010). Ontology and methodology in analytic philosophy. In *Theory and applications of ontology: Philosophical perspectives* (pp. 349-394). Springer, Dordrecht.

Tenove, C. (2020). Protecting democracy from disinformation: Normative threats and policy responses. *The International Journal of Press/Politics*, *25*(3), 517-537.