Science and Engineering Ethics (2023) 29:23
https://doi.org/10.1007/s11948-023-00440-6

ORIGINAL RESEARCH/SCHOLARSHIP

Ethics Inside the Black Box: Integrating Science and Technology Studies into Engineering and Public Policy Curricula

Christopher Lawrence[1] · Sheila Jasanoff[2] · Sam Weiss Evans[2,3] · Keith Raffel[4] · L. Mahadevan[5]

Received: 11 June 2021 / Accepted: 21 April 2023

Abstract
There is growing need for hybrid curricula that integrate constructivist methods from Science and Technology Studies (STS) into both engineering and policy courses at the undergraduate and graduate levels. However, institutional and disciplinary barriers have made implementing such curricula difficult at many institutions. While several programs have recently been launched that mix technical training with consideration of "societal" or "ethical issues," these programs often lack a constructivist element, leaving newly-minted practitioners entering practical fields ill-equipped to unpack the politics of knowledge and technology or engage with skeptical publics. This paper presents a novel format for designing interdisciplinary coursework that combines conceptual content from STS with training in engineering and policy. Courses following this format would ideally be team taught by instructors with advanced training in diverse fields, and hence co-learning between instructors and disciplines is a key element of the format. Several instruments for facilitating both student and instructor collaborative learning are introduced. The format is also designed for versatility: in addition to being adaptable to both technical and policy training environments, topics are modularized around a conceptual core so that issues ranging from biotech to nuclear security can be incorporated to fit programmatic needs and resources.

Keywords Science and technology studies · Complex systems · Pedagogy · Co-production

Introduction

In an era of unprecedented technological change and political uncertainty, modern governance and citizenship demand diverse conceptual tools to make sense of evolving societal challenges. Educators have recognized that traditional university

Extended author information available on the last page of the article

Published online: 22 June 2023

curricula may be too centered within discrete disciplines to instill the hybrid (Jasanoff, 2013; Latour, 1993) proficiencies demanded by many contemporary career paths,[1] and a hodgepodge of programs has emerged to meet those needs by combining technical and humanistic training in various forms. However, these programs often draw their humanistic components from more formal analytic fields—philosophical ethics, neoclassical economics, computational social science—and hence lack the critical and constructivist tools that are well-developed in the field of science and technology studies (STS). While concepts from traditional analytic fields can be useful when dealing with physical or social systems that are stable and well understood, they are often poorly-suited and even counter-productive when applied to many of the complex "wicked problems" (Hoffman, 2020; Rittel & Weber, 1973) that have come to define contemporary public life. When decision-makers come up against radical complexity, discontinuous technological or social change, or intractable political controversies, constructivist skill sets can be crucial for unpacking the politics of knowledge and technology, and for engaging earnestly with skeptical publics (Wynne et al., 1996). And as the complexity (Mitchell, 2009) and interpretive flexibility (Bijker et al., 1987; Collins, 1985) of systems that support human welfare expand and gain recognition, the lack of genuinely hybrid training in these areas may be one of the most consequential gaps in modern higher education.

Meanwhile, STS has brought a decades-long tradition of critical analysis to bear on the complex challenges of modernity.[2] One of the field's distinctive attributes is its application of constructivist analysis to *both* the socio-political *and* the techno-scientific elements of human life (Jasanoff, 2004; Latour, 1993; MacKenzie, 1993; Hacking, 2000; Bijker et al., 1987), and hence its insights are of profound interest to those practical disciplines that endeavor to build, understand and govern advanced technological societies. Yet although STS is well-developed as an academic field, it is not often integrated with the teachings of practical disciplines.[3] While academic STS training is an important enterprise in higher education, there is alongside it a role for more integrated curricula that combine STS training with concepts and methods from those applied disciplines—particularly engineering and policy—that stand to benefit most from STS analysis.

[1] This observation made by authors during conversations with educators at various institutions. See also (National Academy of Science, Engineering and Medicine, 2018).

[2] The STS cannon is exemplified in the Handbooks of STS (Jasanoff, 1995), which have entered their fourth edition (Felt, 2018); flagship journals include *Social Studies of Science* (https://journals.sagepub.com/home/sss), *Science, Technology and Human Values* (https://www.4sonline.org/publications); field fostered by *Society for the Social Studies of Science* (https://www.4sonline.org) which holds annual meetings. While some traditional ethics approaches—such as care-ethics (Giligan, 2008) and virtue-ethics (Carr & Steutel, 1999)—do attend to complexity in ethical life, they are less concerned with how ethical life co-evolves with sociotechnical systems.

[3] A few exceptions: scholarly writing under the heading "engaged STS" generally seeks to promote engagement with practical disciplines, e.g. (Kuhlmann et al., 2017). A "science-outside the lab" approach (Bernstein et al., 2017) incorporates STS but is more explicitly focused on science policy than on breaking open technology. Value-sensitive design (VSD) dovetails with STS in its attention to the values of diverse stakeholder groups (Friedman & Hendry, 2019), but it often formulates a predefined space of possible values that is selected from throughout the evolution of a sociotechnical system (Le Dentac

Toward that end, this article presents a novel framework for course design that can be adapted to a variety of educational settings in schools of practice. The framework is structured around a recurrent thought exercise in which students are asked to consider a scientific or technical artifact that is deemed politically or ethically relevant, and to take apart its technical and political workings in an integrated analysis. This exercise emulates classic STS analyses that *break open the black box* of a technology or scientific claim (e.g.Bijker et al., 1987; Hecht, 2009; MacKenzie, 1993), and it can play two important pedagogical functions. First, it helps instill the confidence (Bandura, 1977) that students need to deconstruct technical or scientific artifacts and map the political assumptions embedded therein. While many institutions require students to take coursework in both technical and humanistic fields, these sensibilities are usually taught separately and treated as distinct "domains." Our pedagogical approach is explicitly designed to integrate them, thereby bringing a critical perspective to bear on supposedly value-neutral activities. Second, as the exercise is carried out recurrently with artifacts drawn from different issue areas,[4] it provides a way to modularize issue content. The list of exercises can therefore be selected to suit the interests of instructors and students, and adapted to programmatic needs. In the experimental course described below, black box exercises were organized into three distinct modules covering biotechnology, information technology and environmental science. Any of these modules could be switched out and replaced with modules on other issue areas, such as nuclear security, pharmaceutical science, epidemiology, or global development.

Threaded through the diverse black-box exercises is a collection of core concepts that form the intellectual backbone of the course design. These were drawn from three theoretical traditions—STS, the study of complex systems (SCS),[5] and critical policy studies (CPS)[6]—and were chosen as points of engagement between STS and the complementary discourses that are already found in engineering and policy schools (respectively). The selected concepts are meant to be concisely presented in a *core concepts crib sheet* during the first week of the course, with each following week featuring one or two of the concepts, backed up for deeper engagement by

Footnote 3 (continued)
et al., 2009; Manders-Huits, 2011). Constructivist STS explicitly attends to the continual coproduction of sociotechnical and value systems, both historically and in the construction of possible future pathways. This is regardless of whether there was a particular attention to values in the processes of design. It thus complements VSD in important ways.

4 This article will make a loose distinction between "conceptual content" that is drawn from theoretical literature, and "issue content"—such as biotechnology, nuclear security, environmental science, information technology, etc.—from which case studies are drawn for exercises, and to which the conceptual content can be applied.

5 The study of complex systems is an interdisciplinary field that investigates physical or social systems that defy traditional reductionist analysis. For a useful summary of the field and its origins for the lay reader, see (Mitchell, 2009). For academic summaries, see Turner and Baker (2016), Rickles (2007) and Astil and Cairney (2015). An example of a prominent journal is *Complex Systems*, https://www.complex-systems.com.

6 Critical policy studies (CPS) generally analyzes policy choices from a post-positivist perspective. For an academic handbook, see (Fischer et al., 2015). An example of a prominent journal in the field is *Critical Policy Studies*, https://www.tandfonline.com/journals/rcps20.

related academic reading. Students are encouraged to draw creatively from the crib sheet concepts throughout the course as they engage with each black-boxed artifact.

One final mechanism for adaptation should be discussed briefly. Students at policy and engineering schools have distinct basic skill sets they need to master, and assignments for courses in this format can be adapted to help develop and integrate with those skills. In policy schools, assignments can be tailored to specific modes of advocacy writing and presentation, such as policy memos and briefings, opinion pieces or policy-relevant research reports. In engineering schools, the conceptual content can instead be incorporated into design projects or technical assessments. Since the black-box exercises are designed to incorporate both technical and interpretive analysis, these components can be weighted and calibrated to match the skill levels and needs that students bring to the course.

In the remainder of this article, we begin by articulating the learning goals that motivate this course design and distinguish it from more traditional attempts to integrate humanistic and technical pedagogy. Next, we exemplify the framework itself by describing an experimental course taught by the authors during the Spring 2020 semester. We list the core concepts that were presented in the crib sheet and the black-box exercises carried out in class. This example is not intended to be strictly replicated, but rather to illustrate the breadth of content that can be brought together into a single coherent course using our design format. We then demonstrate the versatility of the course structure by describing an on-the-fly adaptation that occurred during the COVID-19 crisis, in which students applied the concepts learned in class to analyze rapidly breaking current events. We illustrate student achievements by reporting the research findings of two (anonymized) students who analyzed competing framings of the pandemic in different national settings. The paper concludes with discussion of how this approach might be adapted to other pedagogical settings.

Cultivating the Skills of Deconstruction

Schools of engineering and public policy have long acknowledged reciprocal gaps in their curricula: dearth of ethical or societal consideration in engineering training on one hand (ABET, 2000), and of technical or scientific literacy of future citizens and policy makers on the other. Yet many efforts to fill these gaps retain the conceptual boundaries between disciplines that gave rise to the gaps in the first place. Ethical reflection in engineering schools is often cordoned off in stand-alone courses or extra-curricular activities that are seen as peripheral to the core of engineering training (Hess & Fore, 2018; Sunderland, 2019). Policy schools present the mirror image by offering discrete courses dedicated to various technical problem areas—such as nuclear arms control, cybersecurity, biotechnology, or environment—that are deemed sufficiently important to merit dedicated coursework. Even when cross-domain training is integrated within existing courses, the conceptual separation between scientific/technical and normative/political domains is reinforced by the stated learning goals, which call for some aspects of topics to be "unpacked" for analysis while others are held as fixed and exogenous. For instance, engineering schools cultivate the creative ability to construct and de-construct technical systems to "meet desired human needs…within realistic social, political or ethical constraints" (ABET EC, 2016; Tang & Lee, 2020), where both "needs" and "constraints" are thought to arise outside the system. In the reverse image, policy students are often asked to carry out stakeholder and political analyses that take technical artifacts and scientific knowledge as pre-existing inputs, not as themselves shaped by the value systems within which they come into being.

Yet one of the defining features of modern life is that techno-scientific and socio-political orders are "coproduced" in the sense that each provides the conditions of possibility for, and cannot be understood in isolation from, the other (Jasanoff, 2004; Latour, 1993). Analysis that unpacks these elements *together* is therefore a crucial skill across technical and political fields of practice, but little training in those fields is explicitly designed to integrate them. Such hybrid analysis is also the bread and butter of STS research, which seeks to observe coproduction as it is happening. But most existing disciplinary training in STS is geared toward teaching hybrid analysis as an academic pursuit rather than as a practical skill. The conceptual barrier to hybrid, constructivist training thus remains largely intact in many schools of practice.

The problem with teaching engineering and policy students to take *either* technical *or* humanistic domains as exempt from deconstruction is that real-world technical and political actors grant no such exemptions. A glance at any newspaper front page shows that deconstructing scientific facts and technological artifacts is an everyday occurrence in contemporary public life, whether the issue is disruptive innovation, medical or environmental risk, or a public health crisis (Epstein, 1995; Sarewitz, 2004; Douthat, 2020). Commonly-held normative commitments exhibit similar interpretive flexibility—ethical principles like fairness, privacy, lawfulness and democracy may seem straightforward in the abstract, but they disclose radical ambiguities when interpreted within the divergent digital and bioscientific constitutions of contemporary political cultures. Constructivist skill sets are thus essential to make sense of modern controversies (Collins and Pinch, 1993; Sarewitz, 2004), discern competing and often incommensurable framings (Schon & Rein, 1994) and evaluate the stakes in resolving them.

The goal of our pedagogical approach is to cultivate the confidence and skills to deconstruct techno-scientific artifacts and examine the ethical and political aspects of their inner workings from diverse societal vantage points. This aspiration is distinct from several commonly expressed learning goals—many of them important in their own right—that populate pedagogical literature on engineering ethics. For instance, while instilling ethical principles in aspiring engineers has an important place in their training (Hess & Fore, 2018), our course design seeks to develop students' intuition for how political actors can enact the same ethical principles to support very different technological outcomes, and how technological developments can become intertwined with the emergence of new ethical questions and principles (Jasanoff, 2016). Similarly, while aspiring policy makers should certainly aim to develop various forms of technical and scientific literacy, this course design will also turn their attention to the genealogy (Foucault, 1979) of those literacies, and the ethical commitments embedded therein. It seeks, in other words, to teach technical and political deconstruction as a shared practical skill for aspiring engineers and policy makers.

Course Structure: Conceptual Core and Issue Modularity

This section outlines the primary pedagogical elements of the course—the *core concepts crib sheet* and the *black box exercise*—as they were taught during Spring 2020 semester. Two 75-min sessions were held each week, in a combination of lecture and class discussion. A mixture of theoretical, journalistic and technical readings was assigned in preparation for these sessions (~ 75 pages per week). Mondays were generally devoted to presenting and discussing theoretical readings centered around one or more concepts from the crib sheet, and these concepts were then deployed during the black-box exercises on Wednesdays. Reading responses and advocacy writing assignments were distributed throughout the semester. As the course proceeded, the balance between lecture and class discussion gradually shifted, with instructors taking more of a lead role during early weeks and students gradually becoming more active in the deconstruction exercises. The original plan called for students to be grouped into pairs during the final weeks of the course, and for each group to lead the class through their own black-box exercise to demonstrate how to deconstruct an artifact of their choice (the latter weeks of the course were altered due to Covid-19; see below). This would have constituted the planned final assignment of the course.

The student body for the initial offering of the course was a mix of engineering and non-engineering students, though it was designated as an engineering course. Subsequent years have seen the course move into the general education curriculum, and the crib sheet has been expanded for use in both an STS summer school and for a policy course under development. In addition, this general design is currently being employed by one of the authors in the expansion of attention to the social aspects of engineering within several other required engineering capstone courses.

Presenting Core Concepts in a 'Crib Sheet'

The crib sheet succinctly presented a list of core concepts drawn from the fields of STS, SCS and CPS. Each entry presented a short description of the concept, one or more illustrative examples, and a few assigned and suggested readings to provide a doorway into the relevant literature. All concepts were briefly introduced in lecture during the first week of the course. Later weeks featured one or more concepts for deeper dives and practical application during that week's black-box exercise.

The ultimate learning goal of the crib sheet is to cultivate a general sensibility for the dynamism and constructedness of social, natural and technological orders. Since we envisioned that goal to be pursued iteratively through recurrent exercises in a variety of topical settings, we sought to design a list of concepts that could overlap and combine in diverse ways, and that students could creatively invoke as each exercise unfolded. We also strove to write each entry so that it could provide substantive value upon first read, while also suggesting a clear path for further engagement with additional levels of investment. Finally, concepts were chosen to exploit points of contact and overlap between the three intellectual traditions represented, with a few of the concepts cutting across all three.

In the list of example entries below, we describe our reasons for the selection of each concept. Relevance to applied disciplines is noted, and points of contact between them are indicated. Entries are loosely grouped according to theoretical tradition, though some entries are shared. Definitions as presented in the crib sheet itself are included in Appendix A1.

Framing and incommensurability. The concept of framing appears in various forms throughout the humanities and interpretive social sciences, and it seemed an ideal starting point for the approach taken in this course. The term is commonly used in CPS literature (e.g. Schon & Rein, 1994; Lakoff, 2004), and can be found throughout STS (Epstein, 1996; Jasanoff, 2005). Powerful STS concepts like paradigms (Kuhn, 1962), regimes of perception (Murphy, 2006), techno-political regimes (Hecht, 2009), discourses (Foucault, 1970), and socio-technical imaginaries (Jasanoff and Kim, 2015) also operationalize a similar theme. Throughout the course, students were invited to identify competing ways of framing specific technologies, their underlying metaphors and visions of social good, and points of incommensurability between the frames. They were also asked to strategically deploy and deconstruct frames in brief role-playing simulations, and to consider governance/design challenges that arise when facts emerge from competing frames of meaning.

Boundary work. Boundary work (Gieryn, 1995) is a core term in the STS canon, and refers to the interpretive work done to demarcate bodies of expertise and categories of objects or phenomena. Implicit in this concept is the insight that boundaries between expert and non-expert views can be drawn in multiple ways. Instruction began with Gieryn's four "specimens" of boundary work, and students were thereafter encouraged to identify and track boundary work "moves" carried out by stakeholders as they observed the scientific, technical and political struggles associated with an issue. Alongside this, instructors pointed out the ways that *any* demarcation between categories entails corresponding demarcations of authority and power.

Interpretive flexibility. A common insight of STS analysis is that a single object or artifact can be understood and interpreted in multiple, logically coherent, but often mutually incommensurable ways (Bijker et al., 1987; Collins, 1985). Navigation of interpretive flexibility is not far removed from the imaginative skills commonly utilized in laboratories and technical workplaces, nor from those deployed in effective political or legal decision making. Yet rarely are students on either side of the science-politics divide explicitly taught that the interpretive flexibility of natural, technical and social entities extends "all the way down" (Haslanger, 1995), and hence is subject to continual reexamination and reinterpretation by creative actors.

Performativity of knowledge. Performative concepts of knowledge (Hacking, 1983; MacKenzie, 2009) take knowing not as a passive representation of the world, but as a consequential act that brings the world into being. Bench scientists and engineers may have an intuitive familiarity with this concept, as many of them must routinely consider whether some observed phenomenon is actually the result of their very act of observing. Similarly, big-data practitioners often ask whether their online metrics and algorithms are "accurately measuring" user preferences and behaviors, or are "effectively shaping" them. Policy makers confront similar questions when they write into law various morally significant categories (such as crime, privacy, patient rights, consumer protection, etc.) that then become enacted by citizens to produce unexpected forms of social order or disorder (as when criminalizing drug use promotes drug-related gang behavior and violent crime). These day-to-day practical consequences offer entry points for relating engineering and policy design to STS analyses of performativity.

Construction of subjects and publics. Technologists and policy makers often grapple with the role that their activities can play in shaping human identities, social interests, and political positions and perceptions. Yet many have little awareness of the theoretical traditions that illuminate constructions of identity (e.g. Anderson, 1983; Butler, 1990; Dewey, 1927; Foucault, 1979). Moreover, contemporary debates often make vague reference to a homogenous "public" that is imagined as a pre-existing backdrop for policy or design choices. The black-box exercise format can offer endless teaching opportunities to encourage students to unpack pre-conceived notions of "users," "patients," "viewers," and "publics," and to examine the construction of these groups in relation to the technologies brought to bear on them.

Politics of numbers and probabilities. Technologists often harbor deep concern that quantitative claims have grave social and political consequences, and that they can be "biased" or "manipulated" (e.g. Bergstrom & West, 2020). The STS intervention, however, is to go beyond bias and instill an appreciation that all socially-relevant quantitative systems are constructed (Porter, 1995; Hacking, 1987; Espeland & Sauder, 2007). Several approaches to the deconstruction of numbers were introduced in the black-box exercises. One emulated a "derivation from first principles" format—common to physics and engineering training—that calls upon the student to choose which aspects of a problem are worthy of explicit parametrization in a mathematical representation (e.g. see *Writing down an equation to optimize public welfare* in the appendix). Another focused on the sameness-difference adjudications entailed in any act of counting (e.g., *Is mental illness located in the brain* exercise), which automatically brings boundary work and interpretive flexibility to the fore. Similar deconstructions were carried out with probabilistic claims (e.g., in the latter part of *Anticipating critical transition* exercise).

The above list is deliberately parsimonious, and yet it conveys a broad array of STS sensibilities. Many of them overlap or segue into one another. For instance, *interpretive flexibility* essentially follows from *framing* and *boundary work*, and *construction of subjects* could be presented is an example of *performativity*. These overlaps facilitated continuous transition between concepts and issue areas during class discussion, and offered opportunities for students to make creative connections between them.

The STS concepts above were brought into conversation with some basic concepts from complexity theory. Since the entries listed below are relevant to a broad range of both physical and social systems, they served as touch points for engagement between intellectual traditions, and further segues between issue domains.

Chaotic and complex systems. The physical and social systems that support human welfare—from atmospheres and economies down to social or neural networks—often do not resemble the predictable linear systems that are abstracted in laboratories or analytic mathematics (Rickles, 2007). Cultivating an appreciation of nonlinear systems in their varied forms can alert students to the political challenges and moral hazards of governing and designing technological interventions. A similar appreciation for complexity motivates the STS literature on risk and disaster (Beck and Wehling, 2011; Downer, 2013; Jasanoff & Wynne, 1998). Two variants can serve as starting points: chaotic systems, which display sensitive dependence to initial conditions (Lorenz, 1993); and complex systems, which cannot be reduced to their component parts (Mitchell, 2009).

Critical transitions. A particular challenge that nonlinear systems pose for political and technological governance is that they often exhibit tipping points, where a relatively small stimulus can lead to large, sudden and irreversible changes in the system. In these cases, learning from past experience can be misleading and result in catastrophic choices. Critical transitions can be found in a diverse array of physical and social systems, including climate transitions, economic crashes, and outbreaks of civil unrest (the 2020 BLM demonstrations are a recent example). These are not unlike the abrupt frame shifts in knowledge systems and collective perception that have been the objects of classic STS scholarship. Yet these various tipping points have several common features (Scheffer, 2012) which offer possible pedagogical bridges for shifting across practical issue domains.

Normal versus fat-tailed distributions. Students are often taught to uncritically deploy the normal distribution when making or interpreting statistical claims. Yet the normal distribution relies on assumptions (Reif, 1965, ch. 1) that often are not justified for the complex systems that public officials and engineers must navigate. Many physical and social parameters are better described using fat-tailed distributions, in which extreme events are much more likely to occur than a normal distribution would predict (Taleb, 2007). These extreme events can have outsized influence on the dynamics of societies that are unprepared to deal with such events.

Complex networks. Many consequential features of physical and social systems can be illuminated by graphing their network connections and examining the topological features of those networks (Newman, 2018). For instance, why do wealth and other forms of social power seem to concentrate in the hands of a few individuals? Why do pandemics spread non-uniformly over geographical and social space? Network graphing is a simple exercise from which a broad range of useful intuitions can be gained, and these can be brought into conversation with core STS concerns about power and its accumulation by epistemic and technological means.

The list was rounded out by additional concepts from the field of CPS that were deemed complementary with STS*:*

Collective action problems and tragedy of the commons. The canonical collective action problem succinctly captures a core challenge of governance and the central role of knowledge therein (Axelrod, 1984; Gardiner, 2006; Hardin, 1968). Various articulations were presented, including *tragedy of the commons* and the *prisoner's dilemma*. While these can, at times, appear to oversimplify complex interactions, their common virtue is that they illuminate the role of credible knowledge and trustworthy communication as the essential bedrock of collective action. They thus can serve as teaching tools for conveying the central STS insight that "solutions to the problem of knowledge are at the same time solutions to the problem of social order" (Shapin & Schaffer, 1985).
Panopticism and governance through built environments. Bentham's panopticon, and even Foucault's elaboration of governance through disciplines of optical visibility, are not unfamiliar to undergraduates. They therefore can serve as a gateway for a more general recognition that built infrastructure serves as a form of governance "by other means." Whole traditions of critical social theory flow from that basic insight (Foucault, 1978; Winner, 1986), and these further illuminate how different framings, identities and conceptual boundaries can become "part of the furniture" of everyday social life.
The precautionary principle versus *disruptive innovation*. Cost-benefit analysis is often portrayed as the *sine qua non* of rational decision making. But if the systems underwriting modernity are sufficiently complex or interpretively flexible, it becomes impossible to predict or even define costs and benefits in a persuasive way. Two alternative approaches to rationality have gained traction in recognition of those challenges: the precautionary principle (Kreibel, 2001) and disruptive innovation (Bower, 1995). Fluency in analyzing these competing frames of rationality can help students navigate a wide range of contemporary debates and developments.
Black swans and normal accidents. Policy makers and system designers contending with complex systems must appreciate their propensity to produce extreme events, and how some structural features that herald those propensities can be invisible to traditional probabilistic risk or cost-benefit analyses (Perrow, 1984; Taleb, 2007).

Taken as a whole, this concise list of concepts can develop a critical core of constructivist sensibility that speaks across disciplines. While traditional STS teaching for more academic majors seeks (rightly) to comprehensively cover STS theory and methodology, our approach with the crib sheet is designed to offer more of a foothold that can be strengthened by the students' own creative application of the concepts during the black box exercises. Thus, it is designed to facilitate multiple levels of engagement with each concept, and the concepts chosen should be seen as flexible. Readers should be encouraged to make adjustments to meet needs and interests in their own pedagogical setting and the particular range of topics they are addressing. As we have continued to teach this course, the concepts have shifted each year based on the construction of the teaching team and the goals of the course. For example, the following year, we shifted the course from being an engineering course to meeting requirements for the general education of the entire undergraduate body. Doing so meant we could place more attention on some STS concepts like erasure and commensuration, while still employing technical concepts like fat tails within individual lesson plans.

Exercises: Breaking Open the Black Box

The black-box exercise proceeds by selecting an artefact—a gadget, an infrastructure, an equation, a scientific claim etc.—and deconstructing its technical and political workings in an integrated analysis. The central learning goal is to break down the mutual impenetrabilities of the artifact as a technical, natural and political object. The depth of these deconstructions can be modulated to fit the skills and needs that students bring to the course. For instance, in engineering schools the technical deconstructions can obviously be more advanced—the "calculus version" so to speak—whereas in policy school settings they can instead resemble explanatory journalistic accounts or technical primers. But in either case students should gain some intuition for the complex inner workings of the artifact by employing the crib sheet concepts, while keeping a persistent eye on the political context and consequences of its design. The examples listed below are not necessarily intended to be specifically reproduced so much as to demonstrate the exercise and its versatility. Each entry lists the crib sheet concept(s) featured during that week's introductory (Monday) session (which precedes the exercise on Wednesday), the artifact to be deconstructed, and invitations for student participation. As discussed above, the earliest exercises were mostly driven by the instructors, whereas the later ones invited more active participation from the students. Modules were comprised of three exercises each. Two examples from the module on biotechnology and human health are:

> *Is mental illness located in the brain?* (Featured concepts: *framing and incommensurability*, *politics of numbers*). This exercise drew on work by sociologist Nikolas Rose to unpack common linear explanations of mental illness that seek to isolate neurobiological causes (Rose, 2019). The artifact examined was a set of medical and statistical claims made by public health organizations that allude to a 'global burden of brain disorders' (Wittchen et al., 2011; Collins, 2011; Abbot, 2016; Insel, 2010, 2011). Quantitative claims were deconstructed by examining the diagnostic practices (DSM-5, 2013) that define and delimit "cases" of mental illness to be counted. Empirical practices that render mental illness observable in the brain—including serotonergic interventions (Knutsen et al., 1998; Valenstein, 2005), brain imaging (Insel, 2010) and twin studies (Lahoff, 2010)—were also examined. The instructor began the analysis by identifying the frame that underpins these practices and claims, which imagines mental illness as radiating causally outward from the brain and renders

sources of disorder visible within the brain's biochemical and physiological workings. This frame was contrasted with alternative framings that attribute observable brain changes to social-capital decline or distress "getting under the skin" (Rose, 2019; Monbiot, 2016; Putnam, 1995) or interpret DSM criteria as markers of spiritual "enlightenment" (Obeyesekere, 1985). Incommensurability between these frames is manifest in the divergent ways they lament human tragedy, attribute causal mechanism or responsibility, and prescribe steps for personal and social improvement.

A moratorium on human germ-line gene editing? (*boundary work*). This exercise unpacked scientists' call for a moratorium on human germ-line genome editing (Lander et al., 2019) as a move to position the human womb as a boundary between ethical and unethical applications of CRISPR-Cas9 gene-editing techniques. The proposal was made shortly after Chinese scientist He Jiankui's controversial work that produced the first gene-edited human babies, and banished him from the community of respectable bioscience. As a form of boundary work, it most clearly resembles Gieryn's "expulsion" example, but analogies were made to each of his other three examples (Gieryn, 1995). Students also tracked other mappings in the context of gene-editing debates, including popular distinctions between basic and applied science, medical treatment versus human enhancement, and regulation versus self-government (Jasanoff, 2018). These instances of boundary work were analyzed as efforts to negotiate between popular hopes and anxieties surrounding the discovery and implications CRISPR Cas9.

Module two covered information technology and civic spaces. Two examples:

Is 'privacy' the right frame for addressing the ethics of big data and surveillance? (*panopticism*, *complex networks*, *construction of subjects and publics*). This exercise examined predictive algorithms that rely on very large data sets to identify aggregate statistical correlations. Mechanisms for collecting data from large populations through social media platforms were unpacked. These were then viewed through the frames of individual privacy and contract law. Students were encouraged to identify which ethical issues and mechanisms of political power are rendered visible through each of these frames, and which are rendered invisible. They were then asked to propose alternative frames including ones not resting on concepts of individual rights.

Machine learning and criminal law (*construction of subjects and publics*, *performativity of knowledge*, *complex networks*, *fat-tailed distributions*). The basic artificial neural network for pattern recognition was outlined, along with training techniques and common applications. Special care was taken to illuminate the "vector" structure of input and output data, how the sizes of those vectors depend on the designer's choices in framing the recognition problem, and how those choices in turn filter into the structure of the network itself. The class then examined applications in criminal law enforcement such as planning of police patrol patterns and prediction of flight risk for suspects awaiting trial (Kleinberg, 2016; Barabas, 2020). Students were invited to discuss what types of datapoints might be used in these applications while navigating the distinction between predictive versus criminogenic knowledge and enforcement practices.

Module three examined global environment, inequality and collective action. Two examples:

Anticipating critical transitions (*fat tails*, *critical transitions*, *black swans*, *politics of probability*). Dynamical evolution of a simple climate model (Mahadevan et al., 2010) was taken as the example. Mechanisms for fat-tailed distributions and critical transitions were outlined. The choices embedded in the parametrization of climate were then discussed, along with alternative parametrizations and their implications.
The social cost of climate change (*collective action*, *complex networks*). Climate change was discussed as a problem of "emissions," as a problem of "economic externality" and as a problem of "inequality" (Rayner et al., 1997). Students compared and contrasted the causal mechanisms and interventions suggested by each frame.

The above exercises were designed to be flexible in the sense that the level of technical difficulty could be tailored to skills and needs of the students. In the version of the course reported here, students came from both technical and non-technical majors. We thus used a combination of readings, short lectures and class discussion to explore the inner workings of the selected artifact, rather than mathematical problem sets or laboratory exercises. For instance, for the machine learning exercise, students came to class having read about the use of machine learning in criminal law (e.g. Barabas, 2020) and watched brief tutorials on general pattern-recognition applications (e.g. Sanderson, 2018). The in-class component entailed a brief lecture outlining neural networks, and the exploration of how subjects and publics could be constructed differently through different choices of data structure was carried out in class discussion. Future iterations could be adapted to engineering students by emulating a "laboratory course" approach in which students build basic neural network architectures using Python or MATLAB. In policy school settings, exercises could be adapted into role-playing simulations or technology assessments. In either case, an important learning goal is to *integrate* the technical and humanistic understandings of the artifact's inner workings.[7]

A Real-Time Black-box Exercise: Epidemiological knowledge and Collective Action During the Coronavirus Pandemic

Our institution transitioned to online instruction during March 2020, forcing the last half of the course to be carried out online. This shift, and the pandemic itself, afforded an important moment for the teaching team and the students to return to the purpose of the course and how its design could be modified to capitalize on this new

[7] Much intuition can be gained, for example, from the basic multilayer perceptron (Brownlee, 2016).

learning environment. While the original plan had been for students to pair up and carry out their own black-box exercises on an artifact of their choice, the instructors canceled this final assignment in favor of an exercise focusing on elements of the COVID-19 crisis as they developed in real time. This decision was made in conversation with the students, who recognized that the crisis itself was unfolding as a composite of multiple black-box exercises in real time. Many of the taken-for-granted realities, norms and infrastructures of everyday modern life quickly dissolved before our very eyes, giving the students an unparalleled opportunity to see the inner workings of many of these systems. With examples ranging from epidemiological models and testing practices to critical supply chains and definitions of essential workers, current events offered an unprecedented opportunity for students to deploy crib sheet concepts learned in the course to the events unfolding around them. A collective discussion was carried out to design a new set of group projects to capitalize on these opportunities on the fly.

As a teaching team, our attention to this shift in pedagogical landscape focused on accounting for the student as a whole person undergoing often extreme upheaval. Some of them had family members and close friends hospitalized with COVID-19 while they were still conducting their work for the course, and we employed an open-door model of checking with the students to ensure that their health and well-being were prioritized. We also took the opportunity to engage the students as learning partners (Bovill & Bulley, 2011), working with them in crafting the cases to be studied, and the concepts relevant to analysis.

In the structure that emerged, students grouped themselves around three different themes: quantitative modeling and prediction; institutional practices and governance; and law and equality. Each group carried out a comparative study examining either state- or national-level responses to the COVID-19 crisis, and the diverse constructions of knowledge and politics interwoven with those responses. All student projects were presented in short briefings over Zoom, a video-conferencing platform, on the final day of class, and in written reports published on a course website designed in partnership with the students. Below are two examples of insights that students arrived at during the exercise, both of which demonstrate the students' grasp of the crib sheet concepts and their ability to apply them to real time analysis of events deeply affecting their lives.

> *Epidemiological models, contact tracing and collective action in the United States and South Korea.* The United States relied heavily on quantitative models to inform and justify massive lockdowns that have slowed the spread of COVID-19, but which also threatened the economic livelihood of its citizens. The South Korean government, on the other hand, saw early pushback from the public against the policy of general lockdowns. Instead, it resorted to aggressive testing and contact-tracing measures, which in turn relied on consent and participation from the public on a massive scale. One student examined how the pre-existing political cultures, national experiences and infrastructural preparedness positioned these countries' responses differently though they each identified their first cases of COVID-19 on January 20. Drawing on several core concepts, this student argued that these differences were related, in part,

to divergent *framings* of civil liberties and citizenship, as well as national interests, and resulted in very different modes of *collective action* to mitigate the spread of the virus.

Publics and critical transitions in Germany and Italy. Knowledge-making and collective action in Germany's coronavirus response tend to be more consolidated at the national level, reflecting political culture and institutional precedent for constituting the nation as more communal than factional (Jasanoff, 2005). Italy's response followed a more regional and piecemeal approach, reflecting less institutional capacity to coordinate at the national level, and leading to inter-regional migrations that may have helped to spread the virus. One student, drawing on coursework around the *performativity of knowledge*, tracked how these differences in political institutions became manifest in different ways of tracking infection and mobilizing response. Visual plots of the number of tests performed and known infections (versus time) for each country illustrated how the critical-transition behavior of viral transmission caused relatively subtle differences in response to become dramatically amplified, and produced large differences in ultimate outcomes.

Future Adaptations

The course described above constitutes one iteration of a more general pedagogical approach that combines a conceptual crib sheet with a series of black-box exercises engaging diverse issues at the intersections of science, technology and society. Several dimensions of variation are possible, and we have found that a yearly assessment of the collection of concepts employed meaningfully realigns the course to the learning styles and backgrounds of the students. First, the selection of core concepts can be adjusted to suit the interests and needs of particular mixes and levels of students. Several other relevant concepts include biopolitics (Foucault, 1979), experimenter's regress (Collins, 1985), strange attractors (Lorenz, 1993), and civic epistemologies (Jasanoff, 2005). Similarly, the black-box exercises can be adapted to suit other issue domains, such as nuclear security or forensic science, each of which offers endless artifacts to be deconstructed and brought into conversation with the same crib sheet concepts that we applied in our three modules. The depth of the attempted deconstructions can also be adjusted to suit coursework at both undergraduate and graduate levels, or extend coursework over multiple semesters.

One of our own findings from this offering of this course was that some of the concepts (e.g. panopticism) were employed less frequently in the students' exercises, while others (framing, social construction of publics) featured in most. Systematic analysis by both the students and instructors of the reasons for uptake of some concepts over others could encourage the students themselves to engage in refining the course for future iterations (Cook-Sather et al., 2014). Moreover, a formal assessment of the course, such as that done by Bernstein et al. (2017), could assess the shifts in students' beliefs about the science/society relationship, and the degree to which the crib sheet and black box exercises were seen by the students as contributing to that shift.

The mode of engagement with students can likewise be adjusted to meet diverse pedagogical demands and expectations, as was evidenced through our course shift as a result of the pandemic. While we relied primarily on a combination of lecture and class discussion, many of the black-box exercises naturally lend themselves to the role-playing simulations that are common in policy schools, or to design assignments for applied-science and engineering students. Each type of exercise calls on students to embody different socio-technical vantage points—those of various stakeholders, users, designers, regulators, policy makers, and political strategists—and to strategically deploy competing frames and forms of boundary work to engage (or exclude) particular political and design sensibilities. Finally, the student assignments can encourage a variety of writing genres, ranging from policy memos and briefs to technical assessments and design proposals.

Appendix 1

Core Concepts Crib Sheet.

Eng-Sci 28: Science, Technology and Society.

Spring 2020.

> This course will draw from a set of core concepts to analyze the societal origins and implications of contemporary developments in science and technology. Many of these concepts are drawn from the field of science and technology studies (STS), but other fields, such as complex systems and probability theory, are represented as well. Many of the concepts will be introduced in the first week of class, and following weeks will feature one or two concepts for deeper engagement and academic reading.

Framing and Incommensurability

Below is a picture of the famous Rubin Vase. Many people will look at the picture and see a flower vase. Others will look at the same visual data and see two faces looking at one another. Each is a perfectly valid thing to "see" in the picture, yet it's difficult to see both the vase *and* the two faces at the same time. This is because the two images are *incommensurable* with one another in the sense that the features of a vase cannot *also* be the contours of a face. Cognitive scientists who study the way vision works have pointed to this example to illustrate how the brain can organize visual data in different ways, allowing us to see very different "objects" in the same visual data.

In order to make sense of reality, we must organize our experiences into frameworks of meaning. These interpretive *frames* are needed to separate meaningful data from noise and randomness; identify cause and effect; sort objects into categories; and express or interpret human roles and identities. However, a given reality can give rise to multiple, competing *frames* of meaning that may lead to different understandings. Further, competing *frames* may be *incommensurable* with one another, making it difficult to weigh competing values and sources of evidence against one another. The Rubin Vase example illustrates that even at the level of visual perception, simple tasks like seeing objects in a picture rely on interpretive *framing* to make sense of visual stimuli, and that multiple *incommensurable frames* may lead to seeing different things.

When considering the social implications of science and technology, we will be thinking about *framing* and *incommensurability* at many different levels. Rather than individual tasks of visual perception, we will often consider how groups make sense of things like scientific data, news stories, technological innovations, weather patterns, public health statistics, etc. Yet just as with the Rubin Vase, we will encounter competing interpretive *frames* that allow us to "see" very different objects, risks, choices and facts.

> *Example*—Automobile deaths may be framed as arising from "random accidents," from "careless drunk driving," or from societal preferences for individualized transportation over communal public transit. Each of these

frames reveals different relationships between cause and effect, random and systematic events, personal freedom and determination.

Related reading:

Erving Goffman (1974), *Frame Analysis: an essay on the organization of experience*.

Donald Schon and Martin Rein (1994), Frame Reflection: Toward the Resolution of Intractable Policy Controversies, Ch. 2, "Policy Controversies as Frame Conflicts."

Sheila Jasanoff (2005), *Designs on Nature*, Ch. 2, "Controlling Narratives."

Thomas Kuhn (1962), *The Structure of Scientific Revolutions.*

Social Construction of Knowledge

Our common-sense view of scientific knowledge assumes that seeing or knowing the world is a passive process that is separated from social or political forces. When social forces influence the content of scientific knowledge, we often say that knowledge is "biased." However, scholars who look closely at the construction of scientific knowledge have come to recognize that it is an active process that depends crucially on social factors like trust, authority, shared metaphors and narratives, etc. If social factors are a crucial ingredient in constructing scientific knowledge, then it is meaningless to say that they "bias" that knowledge.

Examples: social construction of vision—We saw in the Rubin-Vase example above that vision is not a passive response to visual stimuli. Rather, our minds must combine that visual stimuli with other, pre-existing concepts that derive from our social experiences, to "construct" an image in our minds. For instance, in order to see a vase, we must have already learned what a vase is. Perhaps we have been given flowers to celebrate a birthday or decorate a new apartment. Someone from another culture who has never had these social experiences would not be able to see a vase in the Rubin-vase picture. Similarly, seeing the picture as two faces talking relies on shared human experiences of conversation. An alien or computer intelligence would scarcely recognize two conversing faces from the picture itself without some prior training.

Related reading:

H.M. Collins (1982), *Sociology of Scientific Knowledge: A Source Book*. Bath, Avon, England: Bath University Press.

Interpretative Flexibility

If you put the concepts of framing and social construction together, it becomes clear that people can see or interpret the world in multiple ways. This includes objects of science and technology, and how they relate to society.

Example: What is a bicycle? It is tempting to think a bicycle as a single object, but it can be many different things at the same time. Different social groups attribute different meanings to the bike because they are trying to address different types of problems.

Social group	What is a bike?	What issue is it involved with?
Campus guard	A menace on footpaths	Campus safety
Professional Cyclist	A means of success	Career advancement
City Planner	A greener alternative to commuting	Sustainable city development

Use this concept often and you will start seeing everything as 'multiple'. This is a fantastic way to start breaking apart arguments about 'technological determinism,' which assumes technology has a single linear path that it inevitably progresses upon. By showing how there are always multiple ways of understanding a piece of knowledge, a technology, or even a 'natural process' in the world, you are demonstrating the constant processes of social construction and maintenance that go on to hold our science and technology together.

Related reading:

H.M. Collins (1981), "Stages in the Empirical Program of Relativism," *Social Studies of Science*, 11(1):3–10 https://doi.org/10.1177/030631278101100101.

T. Pinch et al. (1987), "The Social Construction of Facts and Artifacts," Ch. 1, *The Social Construction of Technological Systems.*

Boundary Work

Where is the border between science and society? How do we assign responsibility for technical or scientific knowledge on one hand, and ethical or political considerations on the other? Sociologists of science and technology have shown that the boundary between science and non-science is established differently in different contexts, through social processes of negotiation, contestation and representation known as *boundary work*. At stake in *boundary work* is the cognitive authority that attends scientific knowledge, and the liability of that knowledge to ethical and political scrutiny.

One common mode of *boundary work* is carried out through specialized technical languages that are comprehensible to "scientists," yet incomprehensible to the "lay people." Another mode can be found in the distinction between "pure" science that claims to seek knowledge for its own sake, and "applied" science that is directed toward serving the goals of society.

Other forms of *boundary work* may involve any contested demarcation between categories: such as "natural/unnatural," "real/fake," "living/nonliving," "human/nonhuman," etc. But when examining these demarcations, we should consider how they may map onto different ways to delegate responsibility, rights and cognitive authority amongst members of a society.

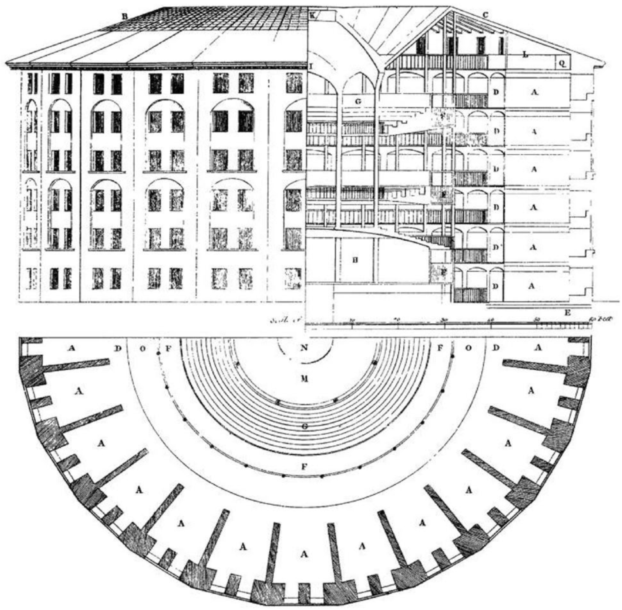
Example—We encountered a stark episode of boundary work during the 2019 public controversy over Dr. He Jiankui's creation of the first CRISPR babies, when professional ethicists (e.g. Dr. Lunshof) and scientists (e.g. Zhang lab statement) called for a moratorium on implanting edited embryos into human wombs. This boundary-work move offers a way to cordon off a space for "basic research" that is allegedly geared toward the pursuit of pure knowledge, and is exempt from the broader ethical and societal concerns that erupted after the Jiankui's controversial experiment.

Related reading:

Thomas Gieryn (1995), "Boundaries of Science," Ch. 18 in *The Handbook of Science and Technology Studies*.

Panopticism

Panopticism is a mode of governance in which political subjects are rendered visible by their built environment, and are thereby encouraged to self-govern. The extreme case is Jeremy Bentham's idealized prison architecture known as the *panopticon* (see Fig. below), which places prison cells in a circular arrangement around a central guard tower. At any given time, prisoners are aware of the *possibility* that they may be observed by a guard in the tower, but unsure whether the guard is looking at that moment. In this way, a strategically designed architecture can act as a force multiplier for the security staff of the prison by instilling self-governance into the prison population itself.



Michel Foucault famously argued that this principle of *governance through visibility* is not limited to prison environments, but has permeated modern societies.

He finds panoptic architecture in seemingly-innocuous environments ranging from military camps to schools, hospitals and workplaces.

Examples—In our contemporary information society, surveillance cameras, big data and machine learning algorithms can function as force multipliers for governance in a way very similar to panoptic architectures that Foucault describes.

Related reading:

Michel Foucault (1978), *Discipline and Punish*, ch. 3, "Panopticism."

Gary Gutting (2005), *Foucault: A Very Short Introduction*, ch. 8, "Crime and Punishment."

Construction of Subjects and Publics

Identity construction is a continual process through which actors' identities are forged, expressed and evolve over time. This process is deeply shaped by experience, which in turn is both socially and technologically mediated. For instance, a given technology or experience may encourage some people to group together as "users," "victims," "consumers," "patients," "experts," "lay-people," "criminals," etc. These identities can, in turn, shape the preferences, political interests, and desires of the people to subscribe to them, or to which they are attached.

Examples—Big data technologies are an important venue for the construction of identities. For instance, in order to record and process data on patient groups for use in medical databases, those groups must be classified based on various markers of identity, such as race, gender, treatment history, etc. But since patients can be classified in an infinite number of ways (see framing and interpretive flexibility), designers must choose which parameters are most "medically meaningful," and these choices in turn have implications for how treatment is prescribed, how patients self-identify, etc. (see performativity below).

Related reading:

Steven Epstein (2007), *Inclusion: The Politics of Difference in Medical Research*, "Introduction: Health Research and the Re-making of Common Sense;" ch. 1, "How to Study a Biopolitical Paradigm;" ch. 2, "Histories of the Human Subject."

Jenny Reardon (2011), "Human Population Genomics and the Dilemma of Difference," in Sheila Jasanoff (edt.), *Reframing Rights*.

Performativity of Knowledge

Performativity of knowledge is the power of knowledge to affect and alter the world.

We often think of knowledge as a "representation" or "mirror" of reality. This is often a useful shorthand, and brings us to ask whether the representation is "accurate" or "biased," and whether the construction of the knowledge is "transparent" or "black-boxed." However, the *representational* concept of knowledge does not account for the fact that, when knowledge is articulated or acted upon, it can feed back and transform the objects or reality that it represents.

Alternatively, we can think of knowledge as *performative* in the sense that it is embodied in human practices and artifacts. Since the *performance* of knowledge takes place in the same reality as the objects that knowledge is "about," we must attend to the interactions through which knowledge can transform the world. The concept of *performativity* can be particularly illuminating when combined with the concept of *framing*.

> *Examples*—To explore the difference between these two concepts of knowledge, consider the music curation algorithms used by Spotify. To first approximation, we might say that the goal of the algorithm is to "discover" the music preferences of the user, and then use that new knowledge of user preference to offer the user listening suggestions that are likely to satisfy those preferences. We might then ask questions about whether that knowledge "accurately represents" the user, or whether it is "biased" toward certain types of music. However, after several listening experiences mediated by the Spotify algorithm, user preference is likely to evolve based on what music is suggested. In other words, Spotify's framing choices for categorizing music taste have the potential to perform or "operate" on the user that is categorized throughout the listening experience, and thus change their musical taste.
>
> Perhaps a more consequential example can be seen in the curation of political discourse via social media's targeted dissemination of news stories. Just as in the Spotify case, this curation of news stories is governed by algorithms that gather knowledge about readers based on their past viewings. A *representational* depiction of the knowledge gathered—which would ask whether an algorithm "accurately discovers" the political views of a user—cannot address the question of how these curation services may have contributed to greater political polarization leading up to the 2016 presidential election.

Related reading:

Donald MacKenzie (2009), *Material Markets*, "Precept 7: Economics *Does* Things," pp. 30–31.

Evan Osnos (2018), "Can Mark Zuckerberg Fix Facebook Before it Breaks Democracy?," *The New Yorker*.

John Austin (1962), *How to Do Things with Words*.

Judith Butler (1990), *Gender Trouble*.

Chaotic and Complex Systems

Many of the physical systems that we model mathematically or encounter in the lab respond to stimuli in predictable ways. With such "well-behaved" systems, we can predict future behavior based on data from previous experiments with the system, or from an analytical understanding of the system's constituent parts. When we measure certain parameters or observables, we can calculate margins of error (i.e. "error bars") to characterize the precision of our measurements, and propagate those through predictive calculations so as to estimate the uncertainty of our predictions. We can also isolate variables such that variation of one doesn't influence our understanding of another. A student may master these basic skills in an undergraduate laboratory course.

However, scientists have come to wonder if the predictability of "well-behaved" systems may be more an achievement of human ingenuity in the lab than a product of how systems behave in the natural world. They have since developed a set of concepts to characterize how predictability of real-world systems can break down, and we will make use of two of these concepts.

The first of these is called *chaos*. A *chaotic* system is one in which very small variation in inputs (or causes) to the system can lead to radically different outcomes. This "sensitive dependence" to initial conditions can be so extreme that even if we can measure inputs with high precision, it may be impossible to calculate probabilities of distinct outcomes. In these cases, our experiences of past outcomes may not help us predict the future, because previous outcomes are very unlikely to recur.

> *Examples*—A classic illustration of chaos is called the "butterfly effect." Imagine that atmospheric weather patterns over the continental U.S. are so unpredictable that a butterfly flapping its wings in San Francisco may create ripple effects that amplify to produce rainstorms in Cambridge. Alternatively, the butterfly may instead sit still, resulting in sunny weather in Cambridge. While this fictitious example is only illustrative, the Earth's atmosphere may indeed respond chaotically to small localized perturbations.

A second concept for characterizing unpredictability is called *complexity*. A *complex system* is one whose component parts may behave in very simple and predictable ways on their own, but when those parts are brought together into a system, their composite behavior becomes very unpredictable. These *complex systems* may produce "emergent phenomena" that cannot be reduced to the behaviors of their component parts.

> *Examples*—Imagine that an individual neuron responds to its inputs in a very predictable way, producing an output pulse if (and only if) its input signals combine to surpass a given threshold. However, when these simple neurons are combined into networks, the aggregate response of the network to input signals can be very difficult to predict or understand.

Related reading:
Edward Lorenz (1993), *The Essence of Chaos*.
James Gleick (1987), *Chaos: The Making of a New Science*.
Dean Rickles et al. (2007), A Simple Guide to Chaos and Complexity, *Journal of Epidemiology and Community Health*.

Critical Transitions

Many complex systems behave unpredictably because they can have tipping points, where some new stimulus causes a sudden and significant change in the state of the system that is irreversible. Many of the systems that humans rely on—such as the climate, economic markets, ecosystems, political systems, the internet, etc.—are complex in this way. Sudden changes in these systems can have catastrophic consequences for humanity.

Examples—The Earth has already measurably warmed due to climate change, yet society has so far been able to manage the consequences. This would seem to suggest that the consequences of further warming will be proportional to what we have already seen, and hence not catastrophic. However, many of the complex systems that society relies on may be approaching a tipping point, at which a small amount of further warming may provoke changes that are qualitatively different from those we have seen in the past. If that is the case, then our past experiences in dealing with climate change may not be a good guide for predicting future consequences.

Related reading:
Marten Scheffer et al. (2012), "Anticipating Critical Transitions."

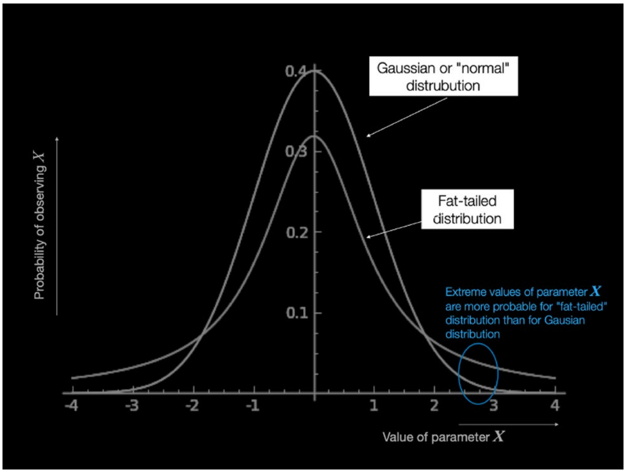
Normal Versus Fat-Tailed Distributions

Many statistical claims about natural or social parameters assume that a parameter follows a *normal* or *Gaussian distribution*, such that we can define an "average" or "norm" value for the parameter, and a "standard deviation" or "variance" from that average value. This turns out to be a good assumption when the parameter in question is itself a composite of multiple statistically-independent random variables. In these cases, it is reasonable for us to expect very large deviations from the average value to be prohibitively unlikely.

For example—imagine I have written a physics article, and I wonder how many times the paper will be cited in the next ten years. I may assume that other physicists each have some probability of reading the article, and readers in turn have some probability of citing it. If the probability of any given physicist having read my paper after ten years is independent of whether others have read it, then I might expect that the number of total citations after ten years might follow a *normal probability distribution.* In this case, I can simply look at the average number of citations for past physics articles, and predict that my article will have a similar citation count in the future.

Unfortunately, many natural and social parameters are made up of composite variables that are not statistically independent, and thus follow a *fat-tailed distribution.* In these cases, probabilistic claims are much more difficult to justify. "Standard deviation" from the average, in particular, becomes meaningless in the case of *fat-tailed distributions*.

For example—if we are interested in the number of citations for my physics paper from the above example, we might suspect that the number of readers in a given year may depend on how many citations the paper might have from previous years. In this case, each citation may enhance the probability that additional physicists may read my paper in the future, and that each reader may in turn cite the paper. If this is so, it may be difficult to predict whether my paper will languish in low readership and citation numbers, or whether citations may "snowball" and result in very large readership.



Related reading:

Nassim Taleb (2007), *The Black Swan: The Impact of the Highly Improbable*, ch. 14: "From Mediocristan to Extremistan and Back"; ch. 15: "The Bell Curve, That Great Intellectual Fraud."

The Politics of Numbers and Probabilities

In order to count the number of instances or members within a given category, we must agree on the definition of the category, as well how to separate distinct instances. If stakeholders agree on these, then quantitative evidence may appear to be "apolitical." However, when stakeholders disagree on how to categorize and differentiate occurrences of a phenomenon—for instance when they *frame* the phenomenon in different ways—then political choices inherent in quantification and mathematics become apparent.

For Example—imagine we want to compare the incidence rates of major depression and anxiety disorders within a population, and that the allocation of national health-care resources depends on this quantitative comparison. It turns out that these two disorders are notoriously difficult to differentiate, making it difficult to determine which diagnosis fits a given patient. Further, even if we agree on how to differentiate the two disorders, it may be difficult to differentiate individual cases from one another. Do we count the number of patients suffering within a population? Or do we count the number of episodes of a disorder, such that a single patient may register several episodes.

Different public-health practitioners may arrive at very different answers for what may seem like a simple question of arithmetic.

Similar considerations pertain to probabilistic claims. In order to write down a probability distribution, we must agree on a space of possible future outcomes, as well as on how that space is delimited or parsed into distinct outcomes.

Tragedy of the Commons and Collective Action

The *tragedy of the commons* is a situation in a shared-resource system where individual users acting independently according to their own self-interest behave contrary to the common good of all users by depleting or spoiling that resource through their collective action. If each member changes their consumption patterns, the resource may be conserved, but that would require *collective action*. Similar to the *prisoner's dilemma*, the *tragedy of the commons* concept helps place individual and collective interests in relation to one another, and illustrate where they may be in opposition. The fact that action must be *collective* in order to overcome these dilemmas is one of the primary challenges of governance.

Related reading:

Garrett Hardin (1968), "Tragedy of the Commons."

Robert Axelrod (1980), "The Evolution of Cooperation."

Herman Daly et al. (2007), "Are We Consuming Too Much — For What?," *Conservation Biology*, 21:5:1359–1362.

The Precautionary Principle

The *precautionary principle* states that if an action or policy has a suspected risk of causing severe harm to the public domain, the action should not be taken in the absence of scientific near-certainty about its safety. Under these conditions, the burden of proof about absence of harm falls on those proposing an action, not those opposing it. This is in contrast to many risk-analytic approaches, in which the burden of proof is to demonstrate the risk outweighs the benefits.

The precautionary principle is intended to deal with uncertainty and risk in cases where the absence of evidence and the incompleteness of scientific knowledge carries profound implications and in the presence of risks of "black swans", unforeseen and un-foreseeable events of extreme consequence (this definition is extracted from Taleb (2014), but modified).

Related reading:

Nasim Taleb et al. (2014), "The Precautionary Principle (with Application to Genetic Modification of Organisms)," *Extreme Risk Initiative, NYU School of Engineering Working Paper Series.*

David Kreibel et al. (2001), "The Precautionary Principle in Environmental Science," *Environmental Health Perspectives*, 109:9.

Additional Black-box Exercises

Gene-drive technology and the re-making of natural and public spaces (*complex systems*, *interpretive flexibility* and *construction of publics*). Gene drive proposals made to the residents of two small islands (Buchthal et al., 2019) were compared as experiments in natural and civic design. The island communities are relatively monocultural and yet distinct, and thus provide good opportunities to examine interpretive flexibility of gene drive technology. Students compared the risks and opportunities of altering the genomes of wild species, and how those were perceived from the perspectives of residents on each island. They were then asked to imagine and evaluate possible mechanisms for containment or propagation of genotypic edits and their social meanings within and beyond island confines, and to deliberate whether the forms of public engagement carried out in these specific locations might be applicable in more general contexts.

Media curation from music to political discourse (*construction of subjects and publics*, *politics of numbers*)? This exercise began by illustrating content-curation algorithms used by Spotify, Youtube and Facebook to suggest content to users. The general concept of the "filter bubble" was discussed, alongside mechanisms of political polarization in the United States. Students then explored how these algorithms allowed Cambridge Analytica and other entities to use "psychographic techniques" to optimize traffic to extreme political content during the run-up to the 2016 presidential election (Osnos, 2018).

Writing down an equation to optimize human welfare (*collective action*, *politics of numbers*). A mathematical expression for inter-temporal welfare was derived from "first principles" (Arrow et al., 2004), and students considered how resources might be distributed across time to "maximize" that expression. They then unpacked those "first principles" and were asked to identify frames, metaphors and ethical values embedded therein. Students were able to quickly identify a frame of "fiscal responsibility" that underlies the mathematics of intertemporal welfare, and they discussed the cultural origins and implications of that ethical sensibility. They were also able to identify several important ethical issues that were washed out of the equation, such as economic inequality across peoples at a given time. The exercise was repeated on alternative frameworks for resource distribution across space and time (Daly et al., 2007).

Declarations

Conflict of interest Not Applicable.

References

Abbot, A. (2016). US mental health chief: Psychiatry must get serious about mathematics. *Nature, 539*, 18–19.

ABET (2000). Improving ethics awareness in higher education. https://www.abet.org/wp-content/uploads/2015/05/Viewpoints_Vol1.pdf Accessed July 10, 2020.
ABET EC (2016). Criteria for accrediting engineering programs. https://www.abet.org/wp-content/uploads/2015/10/E001-16-17-EAC-Criteria-10-20-15.pdf Accessed July 10, 2020.
Anderson, B. (1983). *Imagined communities: Reflections on the origin and spread of nationalism*. Verso.
Arrow, K., et al. (2004). Are we consuming too much? *Journal of Economic Perspectives, 18*(3), 147–172.
Astil, S., & Cairney, P. (2015). Complexity theory and political science: Do new theories require new methods? In *Handbook on complexity and public policy*. Elgar.
Axelrod, R. (1984). *The evolution of cooperation*. Basic Books.
Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84*(2), 191–215.
Barabas, C. (2020). Beyond bias: Reimagining the terms of 'Ethical AI' in criminal law. *12 Georgetown Journal of Law and Critical Race Perspectives*. https://doi.org/10.2139/ssrn.3377921
Beck, U., & Wehling, P. (2011). The politics of non-knowing. In F. Rubio & P. Baert (Eds.), *The politics of knowledge.* Routledge.
Bergstrom, C., & West, J. (2020). *Calling bullshit: The art of skepticism in a data-driven world*. Random House.
Bernstein, M., Reifschneider, K., Bennett, I., & Wetmore, J. (2017). Science outside the lab: Helping graduate students in science and engineering understand the complexities of science policy. *Science and Engineering Ethics, 23*, 861–882.
Bijker, W., Hughes, T. P., & Pinch, T. (1987). *The social construction of technological systems*. MIT Press.
Bovill, C., & Bulley, C. J. (2011). A model of active student participation in curriculum design: exploring desirability and possibility. In C. Rust (Ed.), *Improving student learning (ISL) 18: Global theories and local practices: Institutional, disciplinary and cultural variations* (pp. 176–188). Oxford Brookes University: Oxford Centre for Staff and Learning Development.
Bower, J. L. & Christensen, C. M. (1995). Disruptive technologies: Catching the wave. *Harvard Business Review*.
Brownlee, J. (2016). Crash course on multilayer perceptron neural networks. https://machinelearningmastery.com/neural-networks-crash-course/
Buchthal, J., Evans, S. W., Lunshof, J., et al. (2019). Mice atainst ticks: An experimental community-guided effort to prevent tick-borne disease by altering the shared environment. *Philosophical Transactions of the Royal Society B: Biological Sciences, 374*(1772), 20180105. https://doi.org/10.1098/rstb.2018.0105
Butler, J. (1990). *Gender trouble*. Routledge.
Carr, D. & Steutel, J. (Eds.) (1999). *Virtue ethics and moral education*. Routledge.
Collins, H. (1985). *Changing order: Replication and induction in scientific practice*. University of Chicago Press.
Collins, H., & Pinch, T. (1993). *The golem: What everyone should know about science*. Cambridge University Press.
Collins, P., Patel, V., & Walport, M. (2011). Grand challenges in global mental health. *Nature, 475*, 27–30.
Cook-Sather, A., Bovill, C., & Felten, P. (2014). *Engaging students as partners in learning and teaching: A guide for faculty* (1st ed.). John Wiley & Sons.
Daly, H., et al. (2007). Are we consuming too much—For what? *Conservation Biology, 21*(5), 1359–1362. https://doi.org/10.1111/j.1523-1739.2007.00770.x
Dewey, J. (1927). *The public and its problems*. Henry Holt.
Douthat, R. (2020). In the fog of coronavirus, there are no experts. *The New York Times*.
Downer, J. (2013). Disowning Fukushima: Managing the credibility of nuclear reliability assessment in the wake of disaster. *Regulation and Governance*. https://doi.org/10.1111/rego.12029
DSM-5 (2013). *Diagnostic and statistical manual of mental disorders*. American Psychiatric Association.
Epstein, S. (1995). The construction of lay expertise: AIDS activism and the forging of credibility in the reform of clinical trials. *Science, Technology and Human Values, 20*(4), 408–437. https://doi.org/10.1177/016224399502000402
Epstein, S. (1996). *Impure science: Aids, activism and the politics of knowledge*. University of California Press.

Espeland, W., & Sauder, M. (2007). Rankings and reactivity: How public measures re-create social worlds. *American Journal of Sociology, 113*(1), 1–40.
Felt, E. (2018). *Handbook of science and technology studies*. Fourth Edition. MIT Press.
Fischer, F., Torgerson, D., Durnova, A., & Orsini, M. (2015). *Handbook of critical policy studies*. Elgar Publishing.
Foucault, M. (1970). *The order of things: An archeology of human sciences*. Pantheon Books.
Foucault, M. (1978). *Discipline and punish*. Pantheon Books.
Foucault, M. (1979). *History of sexuality*. Pantheon Books.
Friedman, B., & Hendry, D. (2019). *Value sensitive design: Shaping technology with moral imagination*. MIT Press.
Gardiner, S. (2006). The perfect moral storm. *Environmental Values, 15*, 397–413.
Gieryn, T., et al. (1995). Boundaries of science. In S. Jasanoff (Ed.), *Handbook of science and technology studies.* Sage Publications.
Gilligan, C. (2008). Moral orientation and moral development. In A. Bailey & C. J. Cuomo (Eds.), *The feminist philosophy reader.* McGraw-Hill.
Hacking, I. (1983). *Representing and intervening: Introductory topics in the philosophy of science*. Cambridge University Press.
Hacking, I. (1987). *The taming of chance*. Cambridge University Press.
Hacking, I. (2000). *Social construction of what?* Harvard University Press.
Hardin, G. (1968). Tragedy of the commons. *Science, 162*(5859), 1243–1248.
Haslanger, S. (1995). Ontology and social construction. *Philosophical Topics, 23*(2), 95–125.
Hecht, G. (2009). *The radiance of France: Nuclear power and national identity after WWII*. MIT Press.
Hess, J., & Fore, G. (2018). A systematic literature review of us engineering ethics interventions. *Science and Engineering Ethics, 24*, 551–583. https://doi.org/10.1007/s11948-017-9910-6
Hoffman, M. (2020). Reflective consensus building on wicked problems with the reflect! platform. *Science and Engineering Ethics, 26*, 793–817. https://doi.org/10.1007/s11948-019-00132-0
Insel T (2011). Post by former NIMH Director Thomas Insel: Mental illness defined as disruption in brain circuits. *National Institutes of Mental Health Information Resource Center*, https://www.nimh.nih.gov/about/directors/thomas-insel/blog/2011/mental-illness-defined-as-disruption-in-neural-circuits.shtml.
Insel, T. (2010). Faulty circuits. *Scientific American, 302*(4), 44–52.
Jasanoff, S. (2004). *States of knowledge: The co-production of science and social order*. Routledge.
Jasanoff, S. (2005). *Designs on nature: Science and democracy in Europe and the United States*. Princeton University Press.
Jasanoff, S. (2013). Fields and fallows: The normative logics of STS. In A. Barry & G. Born (Eds.), *Interdisciplinarity: Reconfigurations of the social and natural sciences* (pp. 99–118). Routledge.
Jasanoff, S. (2016). *The ethics of invention: Technology and the human future*. Norton.
Jasanoff, S. (2018). *Can science make sense of life*. Polity Press.
Jasanoff, S., & Kim, S. (2015). *Dreamscapes of modernity: Sociotechnical imaginaries and the fabrication of power*. University of Chicago Press.
Jasanoff, S., & Wynne, B. (1998). Science and decisionmaking. In S. Rayner & E. L. Malone (Eds.), *Human choice and climate change* (pp. 1–87). Battelle Press.
Jasanoff, S. et al. (edt. 1995). *Handbook of science and technology studies*. Revised Edition. Sage Publications.
Kleinberg, J. (2016). A guide to solving social problems with machine learning. *Harvard Business Review*.
Knutsen, B., et al. (1998). Selective alteration of personality and social behavior by serotinergic intervention. *American Journal of Psychiatry, 155*(3), 373–379.
Kreibel, D., et al. (2001). The precautionary principle in environmental science. *Environmental Health Perspectives, 109*(9), 871–876.
Kuhlman, S., et al. (2017). Engaging science technology and policy studies. *EASST Review, 36*, 3.
Kuhn, T. (1962). *The structure of scientific revolutions*. University of Chicago Press.
Lahoff, F. (2010). Overview of the genetics of major depression. *Current Psychiatry Reports, 12*(6), 539–546.
Lakoff, G. (2004). *Don't think of an elephant: Know your values and frame the debate — An essential guide for progressives*. Chelsea Green Publishing.
Lander, E. et al. (2019). Adopt a moratorium on heritable gene editing. *Nature, 567*, 165–168. https://doi.org/10.1038/d41586-019-00726-5
Latour, B. (1993). *We have never been modern*. Harvard University Press.
Le Dantec C., Poole, E., & Wyche, S. (2009). Values as lived experience. In *Proceedings of the 27th international conference on human factors in computing systems (CHI 09),* New York, (p. 1141). ACM Press.
Lorenz, E. (1993). *The essence of chaos*. University of Wisconsin Press.

MacKenzie, D. (1993). *Inventing accuracy: A historical sociology of nuclear missile guidance*. MIT Press.

MacKenzie, D. (2009). *Material markets: How economic agents are constructed*. Oxford University Press.

Mahadevan, L., & Deutch, J. M. (2010). Influence of feedback on the stochastic evolution of simple climate systems. *Proceedings of the Royal Society A, 466*, 993–1003.

Manders-Huits, N. (2011). What is value-centered design? The challenge of incorporating moral values into design. *Science and Engineering Ethics, 17*(2), 271–287.

Mitchell, M. (2009). *Complexity: A guided tour*. Oxford University Press.

Monbiot, G. (2016). Neoliberalism is creating loneliness. That's what's wrenching society apart. *The Guardian*.

Murphy, M. (2006). *Sick building syndrome and the problem of uncertainty: Environmental politics, technoscience, and women workers*. Duke University Press.

National Academy of Sciences, Engineering and Medicine, (2018). Branches from the same tree: The integration of humanities and arts with sciences, engineering and medicine. Consensus Study Report, https://www.nap.edu/read/24988/chapter/1#ix.

Newman, M. (2018). *Networks: An introduction* (2nd ed.). Oxford University Press.

Obeyesekere, G., et al. (1985). Depression, Buddhism, and the work of culture in Sri Lanka. In A. Kleinman (Ed.), *Culture and depression: Studies in the anthropology of cross-cultural psychiatry of affect and disorder.* University of California Press.

Osnos, E. (2018). Can Mark Zuckerberg fix Facebook before it breaks democracy? *The New Yorker*.

Perrow, C. (1984). *Normal accidents: Living with high-risk technologies*. Princeton University Press.

Porter, T. (1995). *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton University Press.

Putnam, R. (1995). Bowling along: America's declining social capital. *Journal of Democracy, 6*(1), 65–78.

Raynor, S. (1997). Zen and the art of climate maintenance. *Nature, 390*(6658), 332–334.

Reif, F. (1965). *Fundamentals of statistical and thermal physics*. Waveland Press.

Rickels, D., Howe, P., & Shiell, A. (2007). A simple guide to complexity and chaos. *Journal of Epidemiology and Community Health, 61*, 933–937.

Rittel, H., & Weber, M. (1973). Dilemmas in a general theory of planning. *Policy Sciences, 4*, 155–169.

Rose, N. (2019). *Our psychiatric future*. Polity Press.

Sanderson, G. (2018). *But what is a neural network? Chapter 1, Deep learning*. https://www.youtube.com/watch?v=aircAruvnKk&list=PLZHQObOWTQDNU6R1_67000Dx_ZCJB-3pi

Scheffer, M., et al. (2012). Anticipating critical transitions. *Science, 338*, 344–348.

Schon, D. A., & Rein, M. (1994). *Frame reflection: Toward the resolution of intractable policy controversies*. Basic Books.

Serewitz, D. (2004). How science makes environmental controversies worse. *Environmental Science and Policy, 7*, 385–403. https://doi.org/10.1016/j.envsci.2004.06.001

Shapin, S., & Schaffer, S. (1985). *Leviathan and the air pump: Hobbes, Boyle, and the experimental life*. Princeton University Press.

Sunderland, M. (2019). Using student engagement to relocate ethics to the core of the engineering curriculum. *Science and Engineering Ethics, 25*, 1771–1788. https://doi.org/10.1007/s11948-013-9444-5

Taleb, N. (2007). *The black swan: The impact of the highly improbable*. Random House.

Tang, B. L., & Lee, J. S. C. (2020). A reflective account of a research ethics course for an interdisciplinary cohort of graduate students. *Science and Engineering Ethics, 26*, 1089–1105. https://doi.org/10.1007/s11948-020-00200-w

Turner, H., & Baker, R. (2016). Complexity theory: An overview with potential applications for the social sciences. *Systems*, *7*, 4.

Valenstein, E. S. (2005). *The war of the soups and the sparks: The discovery of neurotransmitters and the dispute over how nerves communicate*. Columbia University Press.

Winner, L. (1986). *The whale and the reactor: A search for limits in an age of high technology*. University of Chicago Press.

Wittchen, H., et al. (2011). The size and burden of mental disorders and other disorders of the brain. *European Neuropsychopharmacology, 21*, 655–679.

Wynne, B., et al. (1996). Misunderstood misunderstandings: Social identities and public uptake of Science. In B. Wynne (Ed.), *Misunderstanding science? The public reconstruction of science and technology.* Cambridge University Press.

Authors and Affiliations

Christopher Lawrence[1] · Sheila Jasanoff[2] · Sam Weiss Evans[2,3] · Keith Raffel[4] · L. Mahadevan[5]

✉ Christopher Lawrence
Christopher.lawrence@georgetown.edu

Sheila Jasanoff
sjasan@fas.harvard.edu

Sam Weiss Evans
samuel_evans@harvard.edu

Keith Raffel
keith_raffel@harvard.edu

L. Mahadevan
lmahadev@g.harvard.edu

1 Science, Technology and International Affairs Program, Walsh School of Foreign Service, Geogetown University, Washington, USA

2 Program on Science, Technology and Society, Kennedy School of Government, Harvard University, Cambridge, MA, USA

3 Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA

4 Mather House, Harvard University, Cambridge, MA, USA

5 Department of Physics, Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, USA