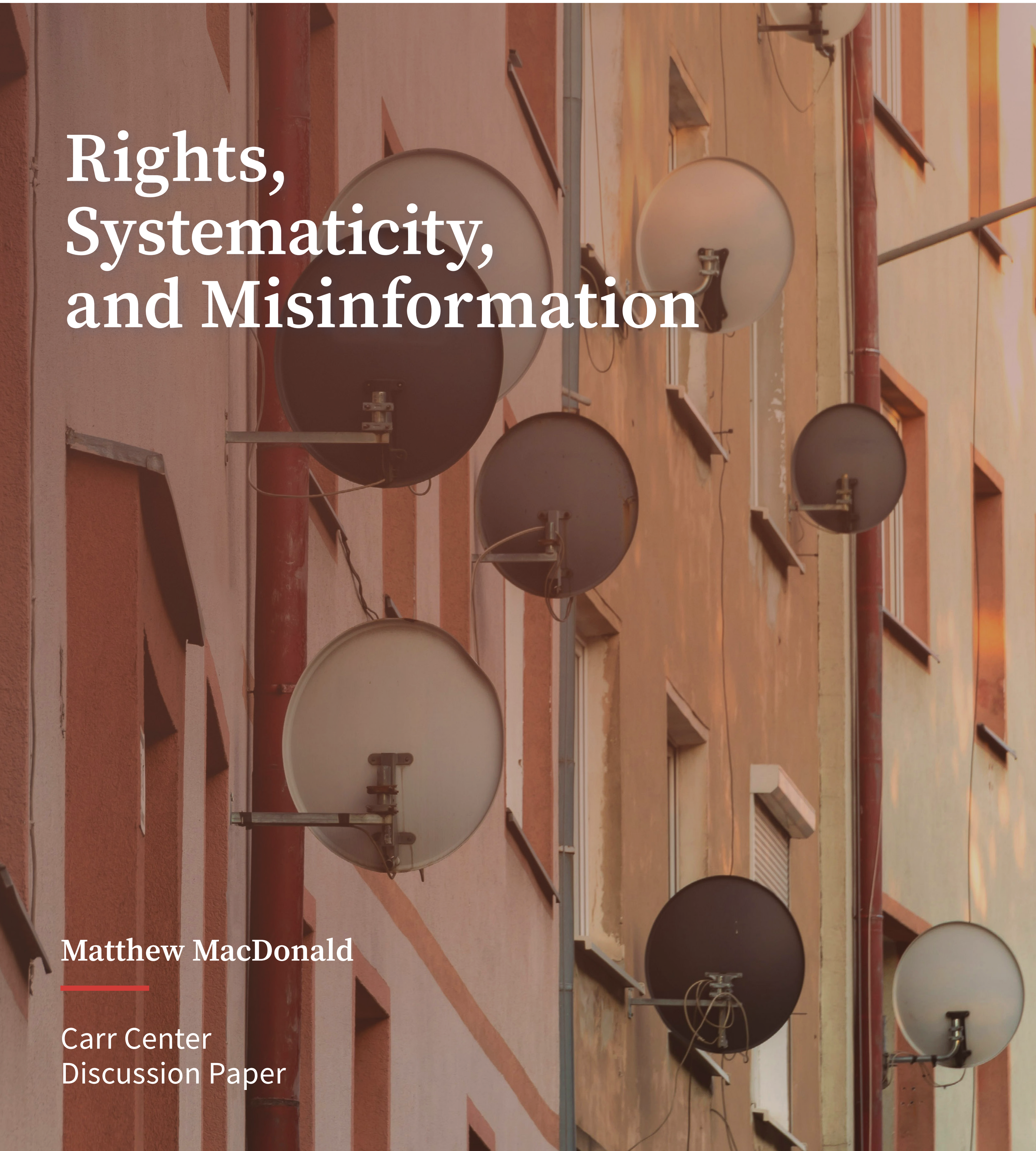
HARVARD Kennedy School
CARR CENTER
for Human Rights Policy

SPRING 2024
ISSUE 2024-03

CARR CENTER FOR HUMAN RIGHTS POLICY
HARVARD KENNEDY SCHOOL



Rights, Systematicity, and Misinformation

Matthew MacDonald

Carr Center
Discussion Paper

Rights, Systematicity, and Misinformation

Carr Center for Human Rights Policy
Harvard Kennedy School, Harvard University
February 27th | Issue 2024-03

Matthew MacDonald
Technology and Human Rights Fellow, Carr Center For Human Rights Policy;
Philosophy Department, Harvard University

The views expressed in the Carr Center Discussion Paper Series are those of the author(s) and do not necessarily reflect those of the John F. Kennedy School of Government or of Harvard University. Faculty Research Working Papers have not undergone formal review and approval. Such papers are included in this series to elicit feedback and to encourage debate on important public policy challenges.

ABSTRACT

The ethical impacts of technology can be understood, in part, by examining how it bears on human rights. Here I offer a general conception of what rights are, which explains their important features and their role as instrumental "mid-points" in moral theorizing. On this conception it is possible for rights to evolve over time, and for new rights to emerge, in response to changes in the "systematicity" of significant hazards. Hateful misinformation is one hazard which is growing increasingly systematic as a result of technologically-driven changes to the infosystem. This motivates carefully re-examining the limits of an established right (the right to free speech), and seriously evaluating the case for a new right (a distinctive right against misinformation).

Communications transmission tower. Image credit: Chris Anderson.

I. Introduction

The growing interweave between online and offline environments has raised ethical concerns about a host of phenomena. These include algorithmic discrimination, violent extremism, public health misinformation, electoral interference, and the erosion of shared public standards for truth. Central to all these concerns is a common set of questions: what is at stake in the development, deployment, and use of the underlying technologies? Are the ethical challenges here qualitatively new, or simply variations on familiar problems? How is the responsibility for addressing them to be distributed between the private sector, governing authorities, and civil society? And what values ought to guide those attempts?

"If technological changes do create new ethical challenges, do some of them require new rights?"

A promising approach to these questions takes human rights as a starting point: whatever else is morally at stake, rights are surely of first importance. This approach is attractive for two reasons: first, human rights are entitlements borne equally and universally, so rights-based theorizing is grounded in an especially deep and inclusive form of moral respect: equal respect for persons as persons. Second, a person's ethical rights have normative bearing on her accompanying institutional rights. For example, although the citizens of different nations hold varying legal entitlements to various kinds of speech protection—different "civic" rights to free speech—those entitlements rest upon a common moral basis—the same "ethical" free speech right—in the sense that latter constrains the range of justifiable variations in the former. These two features—their universality and their normative authority over law—make ethical rights a powerful currency, weighing powerfully (and often decisively) in moral calculation. To say that persons have rights to speak, or vote, or marry whom they wish, is to refer to ethical entitlements which may rarely be abridged.

What is needed for a rights-based framework to be informative, in assessing the ethical impact of new technologies (or new uses for existing ones)? At minimum one needs a taxonomy of which rights persons bear. Established international frameworks (such as the International Bill of Human Rights and the accompanying tradition of international law) are useful here, though we should be careful not to assume that the formal rights they ascribe are strictly identical with the ethical rights taken to underlie them, just as we should be careful to avoid thinking that a person's ethical rights are identical with (or exhausted by) the legal rights granted by her polity. In addition, one needs a method for determining the content of a given right: what *in particular* it guarantees and forbids. If we want to know whether the use of a particular technology "threatens" a right, and what would be required to "protect" it, then we'd better have ways to know what the right contains, especially in new, evolving, or under-theorized situations. It would also be useful to know whether rights can change over time. How easily can a familiar right (like "privacy") evolve in response to technological advances and other permanent developments? And, if technological changes do create new ethical challenges, do some of them require new rights?

In this paper I do two things. First, I set out a general-purpose conception of rights that, I hope, will be helpful for those interested in applying them to the analysis of technology and its uses.[1] Though not a full "theory," it does explain how the content of a right is to be determined, and what I take the ethical function of a scheme of rights to be. On this conception it is possible (but rare) for new rights to emerge, and for their content to evolve.[2] Second, I examine the ways in which technology affects the prevalence of inflammatory and harmful speech. I shall argue that, if it is true that rights can evolve in the ways I describe, there is a case to be made for reexamining an established right (the right to free speech) in view of changes in the global information system, which enable the spread of misinformation, propaganda, and hate speech at greater speed and volume than ever before. These observations also motivate a more radical proposal: that the right to free speech interlocks with a new emerging right against (certain forms of) harmful misinformation.

II. A Proposed Conception of Rights

What I shall present is a conception of "ethical" rights within political morality. These are the enforceable moral entitlements all persons bear equally, distinguishable in principle from the legal rights enshrined in national or international law. One might wonder whether the distinction between ethical and legal rights is coherent. Here is one reason to think that it is: many injustices are best (and most naturally) explained as *gaps* between a per-

[1] Here I follow Rawls and Hart in distinguishing a "concept" from a "conception," where the former re-fers to the outer boundaries of a term's meaningful use, while the latter is an attempt to specify its content more finely. A concept can be the object of multiple reasonable conceptions: here I offer only one that I think is particularly fruitful. John Rawls, *A Theory of Justice* (Cambridge, MA: Harvard University Press, 1971), 5; H.L.A. Hart, *The Concept of Law* (London: Oxford University Press, 1961).

[2] The conception of rights that I present is also compatible with a wide range of background norma-tive and meta-ethical theories about the basis of interests, values, and obligations.

"White people in the US have historically held legal rights to own slaves, but they have never held any such ethical right. This makes sense only if ethical rights are conceptually separable from legal ones."

son's written legal rights and her true underlying ethical rights. Such explanations make sense only if there is, in principle, a meaningful distinction between the two.[3] When members of a minority population are unjustly denied the legal right to vote, or run for office, for example, it makes sense to say that they are denied legal recognition of what is, morally speaking, owed to them. They have ethical rights to participate in self-government even when the legal rights are withheld. The same would be true if, say, some accident of history had resulted in the right to participate in self-government (Article 21) being left out of the UDHR.[4] Even if the right had failed to receive institutional recognition in this form, morally it would be unaffected, for such a right is held by persons whether it is recognized or not.[5] Even clearer cases are those where one's legal rights "go beyond" her ethical rights, granting entitlements that have no underlying moral force. White people in the US have historically held legal rights to own slaves, but they have never held any such ethical right. This makes sense only if ethical rights are conceptually separable from legal ones.[6]

Ethical rights express entitlements—morally authoritative claims to be able to do certain things (e.g. to speak), or to be provided with certain things (e.g. a fair trial), or not be subjected to certain things (e.g. enslavement). Sometimes these entitlements obligate persons directly: for instance, your right to privacy obligates me not to snoop. Other times, the obligations are diffuse, counting for or against broader policies and institutional structures. The right to a fair trial, for example, constrains how a fair judicial system may operate. In those cases, it makes more sense to think of the obligations applying to persons collectively, through their contingent social institutions. What makes a moral entitlement a *right*, I propose, is a combination of both its status alongside other moral elements, and its role in guiding moral judgement, both of which help to determine its content. In particular, a right is a "close-to-all-things-considered" moral entitlement, which:

A) is borne equally by all persons, and directed towards fellow persons (either directly or through collective social institutions); and which

B) expresses necessary conditions for the protection of persons from systematic threats to their shared basic interests, at tolerable cost to those interests, across the full range of reasonably-foreseeable circumstances.

This conception of rights is similar to (but in some ways broader than) the approach taken by Scanlon (1979), who explains that:[7]

> To claim that something is a right, then, is to claim that some limit or requirement on policy decisions is necessary if unacceptable results are to be avoided, and that this particular limit or requirement is a feasible one, that is, that its acceptance provides adequate protection against such results and does so at tolerable cost to other interests. [. . .] What rights there are in a given social setting at a given time depends on which judgments of necessity and feasibility are true at that place and time.[8]

This way of thinking involves a rejection of the view that rights are irreducibly atomic starting points in moral theorizing, subject neither to productive analysis nor revision over time. Instead, I follow Scanlon in thinking that rights occupy a "mid-point" in such theorizing: they are the conclusions of general calculations about the necessity of particular protections against serious standing threats, and the premises of applied casuistry dealing with narrower domains. For this reason, rights take priority over nearly all other normative considerations: they ordinarily function, as

[3] One might think this begs the question against some natural law theorists, who assert that one's legal rights properly include one's ethical rights, even when these differ from the law as it is written. I think this is merely a difference of language: all I want to say is that there can be differences between the rights as written and the rights as they really are, which makes sense only if the latter are not always reducible to the former.

[4] Similarly, although there are disputes about whether Article 16 of the Declaration includes a right to same-sex marriage, a defensible position is that persons have this right (ethically speaking), quite apart from facts about the meaning of the text. United Nations General Assembly, Universal Declaration of Human Rights, Article 16, GA RES/217/A/III (1948), https://www.un.org/en/about-us/universal-declaration-of-human-rights.

[5] I take this to be the point of the US Constitution's 9th Amendment, which states that the legal enu-meration of rights in that document "shall not be construed to deny or disparage others retained by the people." U.S. Constitution, art. 9.

[6] Henceforth, when I refer to "rights" without qualification, I refer to ethical rights.

[7] One way it is broader is that it extends beyond the domain of government policy decisions, concerning also the responsibilities of corporations, non-governmental organizations (NGOs), and related entities. Another difference is that it is grounded in a category of "basic interests," which I shall explain shortly.

[8] Thomas Scanlon, "Freedom of Expression and Categories of Expression," *University of Pittsburgh Law Review* 40, no. 519 (1979), 535-537; Thomas Scanlon, "Rights, Goals, and Fairness," *Erkenntnis* 11, no. 1 (1975), 81-95.

“What makes a moral entitlement a right, I propose, is a combination of both its status alongside other moral elements, and its role in guiding moral judgement, both of which help to determine its content.”

—

Matthew MacDonald
Technology and Human Rights Fellow

Dworkin (1984) says, as "trumps" over other interests in ethical and policy deliberations (indeed, they often preclude a naively consequentialist calculation of those interests), because the most important interests and empirical predictions have already been taken into account when calculating the rights themselves.[9]

This explains both the normative power of rights—the fact that they are generally resistant to tradeoffs in favor of other moral interests—and why such tradeoffs are occasionally permissible nonetheless. For example, it is rarely permissible for a state arbitrarily to curtail citizens' rights to freedom of movement and assembly, by enforcing curfews, breaking up public gatherings, etc. Nonetheless many governments abridged these rights justifiably during the 2020 pandemic lockdowns. The explanation is that the rights, though powerful and usually overriding, are neither absolutely authoritative nor permanently fixed. Rather, they are instrumental safeguards upon an especially deep and important set of moral interests. Though a right may correctly capture what is required to safeguard those interests across the widest possible range of foreseeable contexts, accidents of circumstance can entail that those interests are occasionally better protected in other ways.[10]

The "basic interests" that rights protect require some explanation. The idea is that there are certain things which are reasonably presumed to be important for any person, whatever their other values, goals, or preferences.[11] Very few interests are like this, but those that are should be familiar: they are things like physical safety, economic security, opportunities for self-government, freedom from arbitrary political oppression, some measure of physical health, and so on. These interests can justifiably be ascribed to any person because they are "general-purpose" capacities and preconditions for the pursuit of almost any other goals persons can be expected to have. Whatever one's religious, political, moral, or philosophical convictions and aims, one's instrumental needs will include some familiar elements—elements that are common denominators between otherwise widely varied value systems. Nonetheless the attribution of these interests to any *particular person* is merely "presumptive": it is up to persons themselves to decide whether the interests that can reasonably be ascribed to them are accurate. For instance, a clear-headed person may decide that she cares not a whit for her physical health, and choose a life of smoking, drinking, and binge eating: she may "opt-out" of interests that one would have been justified in attributing to her without prior knowledge. The fact that such ascriptions are presumptive and refutable helps to avoid the worry that a system of rights necessarily involves some objectionable imposition of values upon persons or cultures that might reject them. A presumption can be reasonable even when it turns out to have been incorrect; and one can justifiably make a presumption even when one does not expect that it will prove true in every case. Moreover, basic-interest attributions are already familiar—so familiar that they are almost invisible. When a journalist or human rights monitor reports that the Syrian government bombed its own citizens, for example, they needn't ask whether that contravened the interests of the 11,000 who were killed.[12] Not only can one presume that it did (since one can justifiably ascribe to the victims an interest in physical safety), speculating about whether the presumption is correct would express the opposite of moral respect: it would downplay the horror of that atrocity, putting the focus "in the wrong place." So, there isn't necessarily anything objectionably "impositional" about basic-interest attributions in moral theorizing.[13]

What makes basic interests a morally attractive starting point for theorizing rights? The answer is that they can be justified extremely widely. In addition to simple empirical observations about their near universality (the simple observation that, in fact, people typically value these things), they are instrumentally essential to the vast and varied majority of goals and values that persons and societies hold. This means their moral importance can be (dialectically) justified to the widest possible set of interlocutors—both within each of the widest possible ranges of moral theories and value systems, or as an external neutral set

[9] Ronald Dworkin. "Rights as Trumps," in *Theories of Rights*, ed. Jeremy Waldron (Oxford: Oxford University Press, 1984), 153–67.

[10] One might argue that some rights *are* absolute, for example, those forbidding serious mass atrocities, such as genocide. That view is compatible with my account: all I claim is that *not all* rights have absolute moral authority in every case, because they express close-to-all-things-considered instrumental judgments about what is required to protect the underlying set of basic interests, and such judgments can fall short. Nonetheless it may happen that others perfectly hit the mark: that is, they may be identical to the correct all-things-considered judgments. A right that expressed such a judgement would thus be formulated perfectly, requiring no exceptions to is content or moral force. To allow the possibility of such rights requires no modification of my view.

[11] Compare e.g. Rawls's notion of "primary goods" and Klosko's notion of "presumptive goods." Rawls, *A Theory of Justice*; George Klosko, "Presumptive Benefit, Fairness, and Political Obligation," *Philosophy & Public Affairs* 16, no. 3 (1987), 241-259.

[12] "The Circle of Hell: Barrel Bombs in Aleppo, Syria," Amnesty International (May 18, 2020), https://www.amnesty.org.uk/circle-hell-barrel-bombs-aleppo-syria#.

[13] One might wonder what point there is to a distinction between "rights" and "basic interests": if I'm rejecting the view that rights are a starting point in moral deliberation, and substituting in a deeper set of basic interests, why not simply say that what I've called basic interests are actually what should properly be called our rights? First, I think the distinction is coherent because rights have a more "determinate" character than basic interests—they describe specific enforceable entitlements given what is known about standing threats to the latter, which are more nebulous. Second, I do not claim that basic interests are in fact the starting point for moral deliberation generally: the details will depend on one's background moral theory. Rather, I claim that they are the starting point for theorizing rights, because they occur at a reasonably deep consensus point between different background moral theories.

of considerations between different systems.[14] Thus, whether one seeks to ground a conception of rights within some substantive conception of the good, or in a shared set of "external" moral considerations that is neutral between competing conceptions, one can ground them in basic interests.[15]

III. Systematicity

My basic proposal is thus that rights-claims express moral judgments, calculated at a "close-to-all-things considered" level of generality, about which specific entitlements are necessary for the protection of persons' basic interests from systematic threats. Moreover they must do so at tolerable overall cost to the interests themselves: it must be the case that, on balance, the system of rights does not end up doing more to undermine the interests overall than to protect them.[16]

I now want to turn to the notion of a "systematic threat." Threats to a basic interest are systematic to the extent that they:

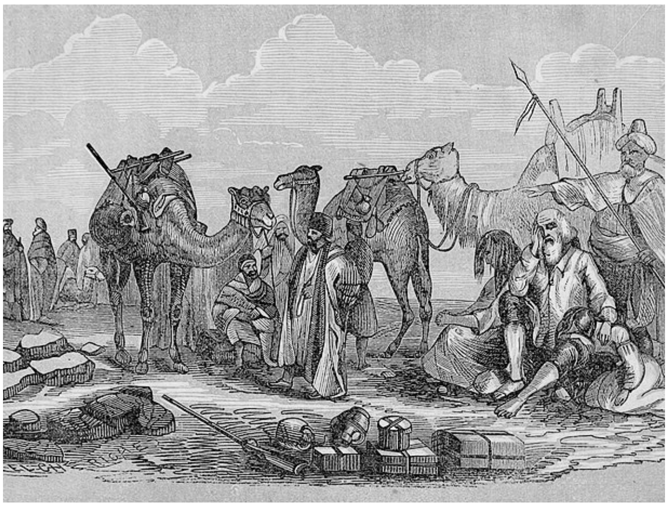
(A) stand to cause significant harm to that interest;

(B) foreseeably recur consistently over time, and across a wide range of societies and circumstances, (e.g. as the result of entrenched human capacities, motivations, or social systems); and

(C) occur in forms that are sufficiently predictable, explicable, and consistent that enforceable political entitlements (i.e. rights) can be formulated to guard against them.

The central idea is hence that rights are protections against well-known major evils. For example, enslavement is one of the most significant and consistently-recurring ways for various basic human interests to be violated, often horrifically. It has emerged, in varied but related forms, across a remarkably wide range of cultures and circumstances, and will continue to do so as long as human societies remain what they are. Thus, it is a systematic threat, significant and consistent enough that an enforceable entitlement (the right not to be enslaved) is required against it.

Some threats may not change much over time: if so, the content of the corresponding rights remains fixed.[17] But the systematicity of other threats can change. This is important for the ethical evaluation of technology, because many such changes are technologically-driven. In the latter part of this paper, I shall consider how changes in the information ecosystem can affect the systematicity of threats, like violent hate speech, propaganda, and misinformation. My general proposal is that changes in a threat's systematicity can alter the content of the corresponding rights.



Engraving of the history of slavery and the slave trade.
Image credit: Internet Archive Book Images.

To see the simplest version of this, consider the impact of new technologies on the right to privacy. At some point in the past, it was impossible to spy on someone secretly through their webcam, simply because there were no webcams. A right to privacy, calculated in say, 1900, could therefore not have included any specific entitlement not to be spied on in this way (though it would have included broader entitlements not to be spied on in general).[18] With the advent of webcams, the

[14] An example of the latter sort of justification would be the ideal of public reason described by John Rawls. John Rawls, *Political Liberalism* (New York: Columbia University Press, 1993); John Rawls, "The Idea of Public Reason Revisited," in *Collected Papers* (Cambridge, MA: Harvard University Press, 1999), 573-615.

[15] This feature also makes the proposed conception of rights (as a component of political morality) consistent with a maximally wide range of background ethical theories, both consequentialist and otherwise.

[16] This condition limits the extent to which basic interests can be traded off against each other, for instance.

[17] This might be one way to explain what are sometimes called "natural rights."

[18] One might think that, as long as the general entitlement entails the specific one, then we needn't include the latter within the content of the right itself, allowing one to resist the idea that rights truly change over time. A reason to think that the content of a right, properly specified, should include at least some of its specific entitlements is as follows. Rights are enforceable entitlements, and effective enforcement requires that the content of what is being enforced can be specified with adequate precision, at least in difficult cases. Doing so requires, as part of the "jurisprudence" of a right, keeping track of which specific entitlement the right's general entitlements entail, and to insist that such entitlements are not, strictly speaking, part of the right itself would be to introduce an unneeded epicycle into our account of rights.

right to privacy expanded to include something new; it continues to do so any time a new form of surveillance emerges.

Less obvious cases are those where new technologies generate entirely new rights. Let us consider two hypotheticals. First, imagine that sometime in the paleolithic past, humans were not yet able to subject each other to deliberate long-term imprisonment (for example, let us imagine that at some very early point, paleolithic humans simply couldn't build structures well enough yet to keep each other forcibly confined). A paleolithic calculation of rights, therefore, could not have identified imprisonment as a systematic threat, and consequently could not have included anything like the right against arbitrary imprisonment that modern humans currently bear (let us ignore for a moment why a paleolithic rights-calculation might have been impossible for other reasons). In the (hypothetical) situation we are imagining, arbitrary imprisonment was not yet a foreseeable systematic threat. This is not to say that there would have been nothing *pro tanto* wrong about maliciously imprisoning one's enemies, if unforeseeable circumstance did occasionally make it possible. My judgment is simply that any such wrongness would have to be analyzed as something other than the violation of a right.

"The systematicity of other threats can change. This is important for the ethical evaluation of technology, because many such changes are technologically-driven."

Compare that example with another: imagine that sometime in the not-too-distant future, technologists invent a dramatic new form of telepathic mind control. Currently we lack that technology, and consequently we lack any distinctive anti-mind-control right. But, just as humans at some point gained a new right against arbitrary imprisonment, perhaps at some point we shall gain a distinctive new right against mind control, if such a technology ever develops.[19]

These thought experiments suggest that it is not particularly surprising or incoherent to think that rights can evolve over time, at least according to the conception I propose. Since they are "mid-points" in moral theorizing, expressing inductive "close-to-all-things-considered" judgments about what is necessary to address long-term threats, they strike a balance between universality and flexibility. Rights can occasionally be recalculated and recalibrated—not often, but often enough—when sufficiently large permanent developments in technology and human social systems alter the overall significance, frequency, and form of those threats (what I have called their "systematicity"). This is because rights are not fixed, unanalyzable moral atoms, but instrumental entitlements contingent partly upon the empirical facts.

While this is not a complete "theory" of rights, I think it is fruitful because it explains some of their important morally-appealing features, and because it explains how new rights (or revisions to existing rights) should be calculated. That calculation would have to involve, first, a description of the basic interests the right is designed to protect; second, a description of which threats to those interests are systematic in the requisite way; third, a description of the specific enforceable entitlements the right contains; and fourth, arguments for both the instrumental necessity of those entitlements, and their tolerability in terms of whatever other basic interests they affect. Because it must yield a close-to-all-things-considered long-term judgement, a rights-calculation might be a substantial task.[20]

The remainder of this paper shall not, therefore, attempt a calculation of that form. Instead, I'll examine how technological changes may affect the structure and content of information environments, both as a general matter and in the particular context of social media. My view is that technologically-induced changes to those environments, interacting with background social conditions, can raise the systematicity of hate speech, propaganda, and misinformation in ways that threaten human lives. The result is that there is a case to be made for the recalculation of an important right, the right to free speech. In addition, there is a growing case to be made for the introduction of a distinctive new right: a right against certain particularly harmful forms of misinformation.

IV. Radio Propaganda

It should be unsurprising that changes in technology can produce changes in a social information environment. Mass communication is, after all, an inherently technological phenomenon, so new technologies can very directly alter which forms of communication are possible within a given infosystem. This extends directly to hate speech, misinformation, and propaganda, which took new forms with the development of new mass communication technologies in the twentieth century.

[19] I have discussed gaining new rights. The question of whether rights can ever be *lost* as the result of permanent developments is interesting, but unfortunately not one that I can pursue here.

[20] Scanlon breaks the calculation down slightly differently, into "(1) ends—the goals or values relative to which the consequences of unfettered discretion are judged to be unacceptable and the constraints proposed are held to be justified; (2) means—the particular constraints that the right in question is taken to involve; and (3) linking empirical beliefs about the consequences of unfettered discretion and about how these consequences would be altered by the constraints the right proposes." Thomas Scanlon, *The Difficulty of Tolerance* (Cambridge: Cambridge University Press, 2003), 152.

One of the most important of these was radio. Notable early exploiters of the medium were the Nazi propagandists, who used it to bring the atmosphere of the mass rally into the homes and workplaces of millions of Germans. Goebbels, in particular, became interested in radio very early, writing in his diary in 1925 (two years after the first domestic broadcasts began in Germany), "Radio! The modern instrument to create philistines! Everything at home! The philistine's ideal!" He was frequently dissatisfied with written propaganda,[21] but called radio "the most important instrument of mass influence that exists anywhere."[22, 23]

RTLM radio station logo. Image credit: Falcon Knight.

From a practical standpoint, radio was easier to control centrally than printed media, and it could react more quickly to unfolding events. Since it was broadcast throughout the day, messages could be tweaked if the initial broadcasts failed to hit the right note. But for Goebbels the important thing was that radio enabled the power of the *voice*, with its nuances of intonation and emotion, and bypassed the more active form of cognitive processing involved in reading. Because hearing is passive and requires less thinking, it was thought to open the mind more easily to suggestion. It could also express something of the sonic atmosphere of the Nazi rallies—always a staple of the party's propaganda—enabling the leader to speak to the listener directly while placing them metaphorically within the mass of hearers. It could be simultaneously very personal and very collective.

Though radio first began to revolutionize mass communication in the 1920's, it remained important throughout the twentieth century, especially in parts of the world with limited rates of literacy or television access. In its 2003 judgment on the role of the hate radio station *Radio Television Libre des Mille Collines* (RTLM) in the 1994 Rwandan Genocide, the International Criminal Tribunal for Rwanda (ICTR) stated that: "The nature of radio transmission made RTLM particularly dangerous and harmful, as did the breadth of its reach […] Unlike print media, radio is immediately present and active. The power of the human voice [adds] a quality and a dimension beyond language to the message conveyed. Radio heightened the sense of fear, the sense of danger and the sense of urgency giving rise to the need for action by listeners."[24, 25]

RTLM's role in the genocide has been much discussed. In a 100-day span between April and July 1994, ethnic Hutu militias, soldiers, and ordinary civilians killed at least 800,000 ethnic Tutsis, moderate Hutu, and Twa (in a total Rwandan population of about 8 million)—a rate of killing four times that of the Holocaust.[26] The killing took place in the context of an ongoing civil war with the Tutsi-led Rwandan Patriotic Front (RPF), and immediately followed the assassination of president Juvénal Habyarimana by unknown attackers.[27] RTLM was closely aligned with the extremist Hutu politicians who seized power immediately after the assassination, and it had already been broadcasting anti-Tutsi hate speech, propaganda, and misinformation for about a year beforehand. It began calling for the extermination of Tutsi civilians as soon as Habyarimana was killed, exhorting all Hutus to participate whilst listing the names and addresses of people to be slaughtered. In their verdict against RTLM's founders, the ICTR stated that: ''The Interahamwe [paramilitary] and other militia listened to RTLM and acted on the information that was broadcast by RTLM. RTLM actively encouraged them to kill, relentlessly sending the message that the Tutsi were the enemy and had to be eliminated once and for all."[28]

RTLM's content before the genocide had been a mix of music, entertainment, and various forms of inflammatory speech.

[21] Keith Somerville, *Radio Propaganda and the Broadcasting of Hatred: Historical Development and Definitions* (London: Palgrave Macmillan, 2012).

[22] David Yanagizawa-Drott, "Propaganda and Conflict: Evidence from the Rwandan Genocide," *The Quarterly Journal of Economics* 129, no. 4, (November 2014), 1947–1994.

[23] Stephen Welch, *The Concept of Political Culture* (London: Palgrave Macmillan, 1993).

[24] Somerville, *Radio Propaganda and the Broadcasting of Hatred*.

[25] International Residual Mechanism for Criminal Tribunals, "The Genocide," United Nations, accessed August 23, 2023, https://unictr.irmct.org/en/genocide.

[26] Somerville.

[27] Habyarimana's plane was shot down during a landing in the capital, Kigali. Scholars are divided over whether the assailants are most likely to have been RPF soldiers, Hutu hardliners, or another group.

[28] Yanagizawa-Drott, "Propaganda and Conflict."

Some of it was misinformation (for instance, alleging that all Tutsi were in league with the RPF; alleging without proof that the RPF had killed Habyarimana; or accusations of cannibalism).[29, 30] Broadcasts also included propaganda (such as extracts from Hutu Power speeches and manifestos) and talk shows hosted by commentators who were popular for their humor, informality, and strong views. Foreign observers sometimes found RTLM disarming: the Canadian Ambassador, Lucie Edwards, stated "There were so many genuinely silly things being said on the station, so many obvious lies, that it was hard to take it seriously. It was like relying on the National Enquirer to determine your policy on outer space."[31, 32], Later, the broadcasts included lists of those to be killed.

It is hard to quantify the direct effect RTLM had upon the genocide. The ICTR refrained from a fine-grained judgement on this question; and, in any case, incitement to genocide is a crime whether it results in genocide or not.[33] Rwanda was simultaneously undergoing both a political crisis and a civil war, so the full causal picture is no doubt complex. Nonetheless there is some research that attempts to isolate the role of radio in particular. In a 2014 study, David Yanagizawa-Drott examined the relationship between the rate of killing and the degree of radio coverage in different areas. Because Rwanda's topography is so varied, and topography affects radio reception, patterns of coverage for RTLM were essentially random and uncorrelated with other determinants of violence. Yanagizawa-Drott found that differences in radio coverage explained about 10% of differences in the rate of civilian participation in the killings, and almost a third of violence by militias.[34] In a subsequent study, Yanagizawa-Drott also found that radio's effect on civilian participation was much stronger in areas with lower rates of access to other forms of media.[35, 36] Another study by Strauss (2019) cautions against exaggerated causal claims, but found nonetheless that there were "statistically significant correlations between radio incitement and higher levels of violence among perpetrators."[37] Putting together quantitative evidence about radio coverage with qualitative evidence (such as interviews with perpetrators), Strauss concludes that there is "some evidence that radio catalyzed the more hard-core among the broader pool of perpetrators."[38]

V. Technological Access

These historical examples illustrate two general points. First, technology that enables a greater speed, volume, and reach of communication can be used for inflammatory purposes; second, inflammatory speech can sometimes affect the actions of hearers, at least in some social conditions. To understand how an infosystem is affected by a particular technology, however, one should not look merely at the technology's "intrinsic" features, or at advances at the "cutting edge" of what is technically possible. Changes in an infosystem can sometimes occur that way—as the result of technological progress simpliciter—but they can also occur as the result of changes in the way that existing technologies are used, such as changes in access.

What made Nazi radio use interesting was not merely the medium's sonic capabilities, or even the content of what was broadcast (such as the use of music and entertainment programs to attract audiences towards Hitler's speeches).[39] It was

[29] Article 19, Broadcasting Genocide: Censorship, Propaganda and State-Sponsored Violence in Rwanda 1990–1994 (London: Article 19, 1996), 112.

[30] Somerville, *Radio Propaganda and the Broadcasting of Hatred*.

[31] Article 19, 84.

[32] Somerville, *Radio Propaganda and the Broadcasting of Hatred*.

[33] The Prosecutor v. Ferdinand Nahimana, Jean-Bosco Barayagwiza, Hassan Ngeze, Judgment and Sentence, Case No. ICTR-99-52-T, ¶ 1015; Susan Benesch, "The Ghost of Causation in International Speech Crime Cases," in *Propaganda, War Crimes Trials and International Law*, ed. Predrag Dojčinovic (New York: Routledge, 2013), 260-274.

[34] One reason for the stronger effect of RTLM broadcasts on militia is that the broadcasts included "organizing" speech (e.g., coordinating which militia groups should go where) in addition to "inciting" speech.

[35] Yanagizawa-Drott, "Propaganda and Conflict."

[36] David Yanagizawa-Drott, "Propaganda vs. Education: A Case Study of Hate Radio in Rwanda," in *The Oxford Handbook of Propaganda Studies*, eds. Jonathan Auerbach and Russ Castronovo (Oxford: Oxford University Press, 2013).

[37] Scott Strauss, "What Is the Relationship Between Hate Radio and Violence? Rethinking Rwanda's 'Radio Machete,'" in *Media and Mass Atrocity: The Rwanda Genocide and Beyond*, ed. Allan Thompson (Waterloo, ON: Center for International Governance Innovation, 2019).

[38] Strauss.

[39] Heidi Tworek, *News from Germany: The Competition to Control World Communications, 1900–1945* (Cambridge, MA: Harvard University Press, 2019).

also the party's aggressive efforts to make sure access to the medium became widespread. One of Goebbels' first priorities as Propaganda Minister from 1933 was to double the number of radio-owners, and that year he ordered the development and mass production of the *volksempfänger* ("people's receiver")—a cheap, state-subsidized radio that cost about half the price of comparable models. It was designed to be sold as widely as possible, and it was heavily promoted towards working-class households. It also had limited range and frequency, so although it could pick up broadcasts by the (Nazi-controlled) RRG network, listeners could rarely hear other stations. The year of its release, the *volksempfänger* accounted for about half of radio sales in Germany; and three-quarters of sales the following year. Over the next five years, radio ownership did roughly double (from 4.3 million in 1933 from to 8.2 million in 1938); it doubled again to more than 16 million by 1941. This part of Goebbels' propaganda effort was a resounding success.[40,41]

Volksempfänger Type D radio, manufacured in 1938.

It is of course impossible to quantify the precise contribution the explosion in radio access made towards the atrocities of the Nazi Regime. However, the example illustrates how technological changes in an infosystem are partly the result of the background political, economic, and social conditions under which a new medium is deployed. Sometimes those conditions affect the pattern of access to it; they can also favor the spread of a medium in somewhat attenuated forms. The volksempfänger was not an innovation at the cutting edge: it was a low-quality version of existing technology, designed to amplify only one kind of message. Not all technological changes to an infosystem involve technological progress.

A similar example is the role of Facebook in the ongoing genocide of Rohingya Muslims in Myanmar. Following the country's haphazard transition towards democracy from 2010-15, and with few markets left to conquer elsewhere in the world, Facebook began partnering with cellphone companies to expand internet access in Myanmar from 2013 onwards. This was done by making cheaper SIM cards and smartphones available, dropping the price of the average SIM from more than $3000 to less than $1.[42] The smartphones were pre-loaded with a readily accessible, Facebook-centric form of internet called "Free Basics" (also known as "Internet.org"), which allowed users who signed up to Facebook to view a stripped-down version of the site (and a few other sites) without data charges.[43] According to Alan Davis, a researcher who visited Myanmar at the time, the effect was that "[f]or the vast majority of Myanmar citizens and probably near enough all of Myanmar youth, Facebook *is* the internet—and the internet is Facebook. People do not have the time, or the eyes, to visit websites—not when you have to squeeze a website so it looks good on a one-inch square screen. Facebook and social media therefore started to become the dominant factor in Myanmar communication."[44]

Facebook was apparently either unaware or indifferent to the potential for misuse. Adhering to its "move fast and break things" mantra, the company aggressively pursued growth without regard for the risks of applying its formula to a country without strong traditions of democracy or civil society, and in which it had neither a physical presence nor particular expertise (for example, the company employed only one Burmese speaker, based in Dublin, to moderate content for a country of 50 million).[45,46] The growth was certainly dramatic: the proportion of

[40] Ernst Kris and Hans Speier, *German Radio Propaganda* (New York: Oxford University Press, 1944), 51.

[41] "German Radio: The People's Receiver," U.S. Holocaust Memorial Museum, accessed August 17, 2023, https://exhibitions.ushmm.org/propaganda/german-radio-the-peoples-receiver.

[42] Sam Petulla, "This SIM Card Used to Cost $3,000. Democracy May Bring It Down to Zero," *Quartz* (March 14, 2013), https://qz.com/62523/this-sim-card-used-to-cost-3000-democracy-may-bring-it-down-to-zero.

[43] Stephanie MacLellan, "Fake News, Dangerous Speech and Mass Violence: Challenges for Social Media in the Developing World," in *Media and Mass Atrocity: The Rwanda Genocide and Beyond*, ed. Allan Thompson (Waterloo, ON: Center for International Governance Innovation, 2019).

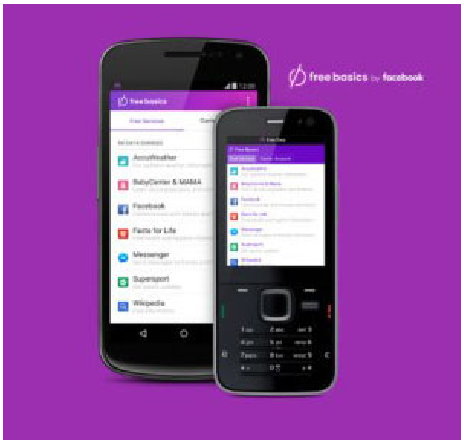
[44] Alan Davis, "Hate Speech in Burma," in Media and Mass Atrocity: The Rwanda Genocide and Beyond, ed. Allan Thompson (Waterloo, ON: Center for International Governance Innovation, 2019).

[45] Mark Latonero and Aaina Agarwal, "Human Rights Impact Assessments for AI: Learning from Facebook's Failure in Myanmar," Carr Center for Human Rights Policy (2021), https://carrcenter.hks.harvard.edu/files/cchr/files/210318-facebook-failure-in-myanmar.pdf.

[46] Laurie Kim, "Tech Accountability in Face of Genocide: Gambia v. Facebook," Emory International Law Review 36, no. 1 (2022), https://scholarlycommons.law.emory.edu/eilr/vol36/iss1/6.

Myanmar’s population with internet access rose from 1% under the pre-transition military regime, to more than 50% in 2015.[47,48] But it was characterized by an explosion of misinformation and hate speech, as a population almost entirely without internet literacy rapidly entered an infosystem dominated by engagement algorithms, filter bubbles, unverified sources, and access to information at greater speeds and volumes than ever before.

What was not obvious to users was that many of the telecommunications networks were still controlled by the country’s military, as were many of the groups and pages that dominated Burmese-speaking Facebook feeds.[49,50] These apparently included innocuous-looking groups, fake news media accounts, and pages devoted to Myanmar’s celebrities, which started out posting ordinary content before pivoting to hate speech.[51] Social media was also exploited by hardliner Buddhist groups, such as the radical nationalist *Ma Ba Tha* movement, to spread atrocity misinformation against Muslim minorities in the country. These included rumors that Muslims were plotting “to take over Myanmar by having too many children and converting innocent Buddhists,”[52] or rumors that Muslims were stockpiling weapons and explosives in mosques in preparation for widespread terrorist action.[53,54]



Advertisement for phones with 'Free Basics'

The rapid increase in the speed and volume of misinformation, rumors, and sheer hate speech was accompanied by a rise in anti-Muslim violence.[55,56] For example, in 2014, two people were killed and about twenty injured during riots in the city of Mandalay following false rumors on Facebook that a Muslim tea vendor had raped a Buddhist woman.[57,58] In mid-2016, Buddhist mobs forced dozens of Muslims to flee their village in the Bago region after false rumors about plans to build a mosque there.[59] These events were precursors to the mass crackdown by military forces, aided by Buddhist civilian groups, on Rohingya Muslims in Rakhine state in 2017. Over the first month, nearly 300 villages were razed, killing nearly 7,000 and displacing about 725,000 refugees, with many fleeing over the border into Bangla-

[47] Latonero & Agarwal, “Human Rights Impact Assessments for AI.”

[48] Freedom House, “Burma,” in Freedom on the Net 2012: A Global Assessment of Internet and Digital Media, ed. Sanja Kelly, Sarah Cook, and Mai Truong (New York: Freedom House, 2012), https://freedomhouse.org/sites/default/files/Burma%202012.pdf.

[49] Davis, “Hate Speech in Burma.”

[50] Latonero & Agarwal, “Human Rights Impact Assessments for AI.”

[51] Paul Mozur, “A Genocide Incited on Facebook, With Posts From Myanmar’s Military,” *New York Times* (October 15, 2018), https://nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html.

[52] Paul Mozur, “A Genocide Incited on Facebook, With Posts From Myanmar’s Military,” *New York Times* (October 15, 2018), https://nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html.

[53] Sheera Frenkel, “This Is What Happens When Millions of People Suddenly Get the Internet,” *Buzzfeed* (November 20, 2016), www.buzzfeed.com/sheerafrenkel/fake-news-spreads-trump-around-the-world?utm_term=.doK0VQjAY#.lyK1E0qgv; Nick Baker, “How Social Media Became Myanmar’s Hate Speech Megaphone,” Myanmar Times (August 5, 2016), www.mmtimes.com/national-news/21787-how-social-media-became-myanmar-s-hate-speech-megaphone.html.

[54] MacLellan, “Fake News, Dangerous Speech and Mass Violence.”

[55] MacLellan.

[56] Baker, “How Social Media Became Myanmar’s Hate Speech Megaphone.”

[57] Mark Frohardt and Paula Orlando, “The Role of the Media in Fostering a Culture of Critical Engagement in the Context of Mass Atrocities: Examples from Rwanda, Colombia and South Sudan,” in *Media and Mass Atrocity: The Rwanda Genocide and Beyond*, ed. Allan Thompson (Waterloo, ON: Center for International Governance Innovation, 2019).

[58] Davis, “Hate Speech in Burma.”

[59] MacLellan, “Fake News, Dangerous Speech and Mass Violence.”

desh.[60] The attacks also involved mass executions, rapes, and torture. One soldier testified that they were instructed to "[k]ill all you see, whether children or adults."[61]

In 2018, the UN's Independent International Fact-Finding Mission in Myanmar ("IIFFM") issued its judgment that the atrocities constituted genocide against the Rohingya. Marzuki Darusman, the Mission's chairman, stated that misleading and hateful posts on Facebook had played a "determining role" in the violence,[62] and Yanghee Lee, Special Rapporteur on Human Rights in Myanmar added that "Facebook has now turned into a beast."[63] The IIFFM's official report (issued in September 2018) extensively cites Facebook posts by public figures in Myanmar in its judgment that the platform contributed significantly to the dissemination of hate speech and governmental propaganda, in which Rohingya were inhumanely demonized and portrayed as an existential threat, and details how Facebook was used to "exacerbate a climate in which hate speech thrives and in which individuals and groups may be more receptive to calls of incitement to violence."[64] It also describes the prevalence of "deeply exclusionary and dehumanizing" online hate speech towards the Rohingya, and notes that is impact is "compounded by the stream of false or incomplete information," concluding that "[t]he role of social media is significant. Facebook has been a useful instrument for those seeking to spread hate, in a context where for most users Facebook is the Internet."[65,66]

The IIFFM also concluded that Facebook's efforts to stem further escalations were "slow and ineffective." From 2013 until at least six months after the genocide began, Facebook consistently ignored warnings from civil society groups of the surge in violence and online misinformation, including an open letter from human rights monitors that noted the connections between specific online rumors and subsequent offline attacks. It also hampered efforts by independent researchers to monitor its practices of data-collection and content-recommendation, and it removed specialized functions such as "Graph Search" that human rights monitors had relied upon to connect specific military groups with the crackdowns.[67, 68] It was not until well into the genocide that Facebook began removing accounts controlled by Myanmar's military and retiring its Free Basics service, and it remained resistant to compliance with the ongoing Independent Investigative Mechanism for Myanmar, at least until it began to face substantial legal pressure and media attention.[69]

The spread of violent content was closely tied to Facebook's business model. In an independent report, Amnesty International concluded that "[Facebook's parent company] Meta's contribution was not merely that of a passive and neutral platform that responded inadequately in the face of an unprecedented crisis. In reality, Meta's content-shaping algorithms proactively amplified and promoted content on the Facebook platform which incited violence, hatred, and discrimination against the Rohingya."[70]

[60] Office of the UN High Commissioner for Human Rights, "Report of the Independent International Fact-Finding Mission on Myanmar," Human Rights Council (September 12, 2018), www.ohchr.org/Documents/HRBodies/HRCouncil/FFM-Myanmar/A_HRC_39_64.pdf.

[61] Hannah Beech, Saw Nang, and Marlise Simons, "'Kill All You See': In a First, Myanmar Soldiers Tell of Rohingya Slaughter," *New York Times* (September 8, 2020), https://www.nytimes.com/2020/09/08/world/asia/myanmar-rohingya-genocide.html.

[62] Beech, Nang, and Simons.

[63] Tom Miles, "U.N. Investigators Cite Facebook Role in Myanmar Crisis," *Reuters* (March 12, 2018), https://www.reuters.com/article/us-myanmar-rohingya-facebook/u-n-investigators-cite-facebook-role-in-myanmar-crisis-idUSKCN1GO2PN.

[64] Kim, "Tech Accountability in Face of Genocide."

[65] MacLellan, "Fake News, Dangerous Speech and Mass Violence."

[66] Office of the UN High Commissioner for Human Rights, "Report of the Independent International Fact-Finding Mission on Myanmar."

[67] Kim, "Tech Accountability in Face of Genocide."

[68] Latonero & Agarwal, "Human Rights Impact Assessments for AI."

[69] Kim, "Tech Accountability in Face of Genocide."

[70] Amnesty International, "Myanmar: The Social Atrocity: Meta and the Right to Remedy for the Rohingya" (September 29, 2022), https://www.amnesty.org/en/documents/asa16/5933/2022/en/.

Internal communications revealed that the company was aware that its recommendation algorithm and business model—which promotes viral and engaging content, no matter how inflammatory—were contributing to offline harm. Internal research prior to the genocide concluded that "[o]ur recommendation systems grow the problem."[71] A further internal document in 2019 noted: "We have evidence from a variety of sources that hate speech, divisive political speech, and misinformation on Facebook and the family of apps are affecting societies around the world. We also have compelling evidence that our core product mechanics, such as virality, recommendations, and optimizing for engagement, are a significant part of why these types of speech flourish on the platform."[72]

VI. Social Media

The Myanmar example illustrates that rapid technological changes in an infosystem can, under some conditions, greatly amplify inflammatory speech. These conditions are not limited to Myanmar: as human rights monitors and social media companies themselves have noted, developing countries around the world have seen hate speech and atrocity misinformation explode within their information environments. Anti-Muslim violence in India,[73] ethnic conflicts in South Sudan,[74] and fighting within the Tigray and Oromo regions of Ethiopia[75] have all been associated with violent social media content, in ways that are similar to what happened in Myanmar.

These changes are driven in part by the intrinsic features of the technologies themselves: that is, the ability of content on social media to spread far more widely, quickly, and at much greater volumes than ever before, resulting in infosystems that can easily become over-saturated.[76] Within these environments it can be difficult to identify reliable sources, especially in countries with relatively weak civil society institutions, low internet literacy, poor education, and histories of top-down information control.[77] The difficulty of identifying reliable sources can sometimes motivate users to rely on posts by their family and friends. But, even in those cases, well-meaning users can end up spreading false rumors and fearmongering without double-checking their reliability. For instance, rural India has experienced outbreaks of mob violence following rumors on WhatsApp accusing minorities of rape, murder, and child kidnapping.[78] According to Indian digital expert Durga Raghunath "[m]any of the issues people see on these platforms have an emotional connect, and because the information comes to us via family and friends, the inclination to double check is very low."[79] The fact that WhatsApp is encrypted has also made it harder for human rights monitors to spot dangerous misinformation quickly enough on the platform to sound the alarm.[80]

To some extent social media has democratized news content, allowing anyone to communicate with millions of others. But this feature has also given voice to extremists, trolls, governmental propagandists, scammers, and sheer idiocy. The centralization of news into social media platforms has made these

[71] Cited by Amnesty International, "Myanmar: The Social Atrocity."

[72] Amnesty International.

[73] Chinmayi Arun and Nakul Nayak, "Preliminary Findings on Online Hate Speech and the Law in India," Berkman Klein Center (December 9, 2016), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2882238.

[74] PeaceTech Lab, "Social Media and Conflict in South Sudan: A Lexicon of Hate Speech Terms" (2017) www.peacetechlab.org/s/PeaceTech-Lab_-SouthSudanLexicon.pdf.

[75] Peter Mwai, "Ethiopia's Tigray Conflict Sparks Spread of Misinformation," *BBC* (November 11, 2020), https://www.bbc.com/news/world-africa-54888234; Jack Burnham, "From the Internet to Ashes: Disinformation and the Tigray War," NATO Association of Canada (September 8, 2022), https://natoassociation.ca/from-the-internet-to-ashes-disinformation-and-the-tigray-war/; Lee Hale and Eyder Peralta, "Social Media Misinformation Stokes A Worsening Civil War in Ethiopia," *NPR* (October 15, 2021), https://www.npr.org/2021/10/15/1046106922/social-media-misinformation-stokes-a-worsening-civil-war-in-ethiopia; Claire Wilmot, Ellen Tveteraas, and Alexi Drew, "Dueling Information Campaigns: The War Over the Narrative in Tigray," The Media Manipulation Casebook (Aug. 20, 2021), https://mediamanipulation.org/case-studies/dueling-information-campaigns-war-over-narrative-tigray; Kaamil Ahmed, "Ethiopia Suspends Aid Groups for 'Spreading Misinformation,'" *The Guardian* (August 6, 2021), https://www.theguardian.com/global-development/2021/aug/06/ethiopia-suspends-aid-groups-for-spreading-misinformation.

[76] Simon Cottle, "Beyond Rwanda? Reporting Atrocity in a Changing Communications Environment," in *Media and Mass Atrocity: The Rwanda Genocide and Beyond*, ed. Allan Thompson (Waterloo, ON: Center for International Governance Innovation, 2019).

[77] MacLellan, "Fake News, Dangerous Speech and Mass Violence."

[78] Nadim Asrar, "In India, WhatsApp Stirs Up Deadly Rumours," *Al Jazeera* (July 17, 2018), www.aljazeera.com/indepth/features/india-whatsapp-stirs-deadly-rumours-180717073333535.html.

[79] Ayeshea Perera, "The People Trying to Fight Fake News in India," *BBC* (July 24, 2017), https://www.bbc.com/news/world-asia-india-40657074.

[80] MacLellan, "Fake News, Dangerous Speech and Mass Violence."

platforms a primary way for politicians to communicate with their audiences, and news journalists to pick up content that is subsequently repeated in other forms.[81] And the global reach of these platforms has allowed inflammatory speech to cross borders more easily. This includes speech that aims at election interference, such as the activities of the Russian Internet Research Agency propaganda unit, or the Israeli cyber contractor Team Jorge, which has interfered in elections across Africa.[82] It also includes speech that aims to incite violence elsewhere, such as inflammatory speech in South Sudan's infosystem, circulated by members of the country's diaspora.[83, 84, 85, 86]

As we saw in the case of Myanmar, the business model of platforms like Facebook, Twitter, YouTube, and WhatsApp essentially amplifies these phenomena. Social media companies have evolved rapidly and continuously, and expanded quickly into many different countries, disrupting infosystems in ways that can be hard for regulators and human rights monitors to address quickly. Because they are competitive entities, the platforms are notoriously reticent to share their data and algorithmic designs with outside researchers, making it hard to track their long-term effects; and further incentives against transparency arise from the threat of legal culpability. At the same time, social media platforms have become increasingly thorough at collecting user data, which is used to tailor each user's content feed to maximize their engagement. The core business model involves keeping users online as long as possible by promoting the most engaging content, allowing the platform to sell more ads. Yet both independent empirical studies and internal reports by the companies themselves have shown that the most engaging forms of content are frequently false, anger-inducing, and polarizing, such as calls for violence and demonizing accusations against a target group.[87] These infosystems optimize for outrage, not truth.

VII. Re-Examining Rights

My view is that social media is making various forms of harmful speech increasingly systematic. I said earlier that a threat is systematic to the extent that it recurs repeatedly and at sufficient scale to harm basic interests such as physical safety and the preservation of human life, and that it occurs in forms that are sufficiently explicable that enforceable entitlements (that is, rights) can be formulated against it. Atrocity misinformation, accompanied by hate speech and calls for violence against minority groups, is a threat to the lives and safety of the target groups, at least under certain social conditions or in combination with other causes. This general point was illustrated by the examples of Rwanda and Myanmar, though one could also look to the role of Serbian media during the Bosnian Genocide of 1995, the online rumours that incite mob violence in India, or the contribution of misinformation to the conflict

[81] Jennifer Forestal, "Beyond Gatekeeping: Propaganda, Democracy, and the Organization of Digital Publics," The Journal of Politics 83, no. 1 (2020): 306-320.

[82] Stephanie Kirchgaessner and Jason Burke, "Political Aides Hacked by 'Team Jorge' in Run-Up to Kenyan Election," *The Guardian* (February 15, 2023), https://www.theguardian.com/world/2023/feb/15/political-aides-hacked-by-team-jorge-in-run-up-to-kenyan-election; Manisha Ganguly, "'Aims': The Software for Hire That Can Control 30,000 Fake Online Profiles," *The Guardian* (February 14, 2023), https://www.theguardian.com/world/2023/feb/15/aims-software-avatars-team-jorge-disinformation-fake-profiles; Stephanie Kirchgaessner et al., "Dark Arts of Politics: How 'Team Jorge' and Cambridge Analytica Meddled in Nigerian Election," *The Guardian* (February 16, 2023), https://www.theguardian.com/world/2023/feb/16/team-jorge-and-cambridge-analytica-meddled-in-nigeria-election-emails-reveal; Stephanie Kirchgaessner et al., "Revealed: The Hacking and Disinformation Team Meddling in Elections," *The Guardian* (February 14, 2023), https://www.theguardian.com/world/2023/feb/15/revealed-disinformation-team-jorge-claim-meddling-elections-tal-hanan.

[83] MacLellan, "Fake News, Dangerous Speech and Mass Violence."

[84] PeaceTech Lab, "Social Media and Conflict in South Sudan."

[85] MacLellan, "Fake News, Dangerous Speech and Mass Violence."

[86] Geoffrey York, "Social Media in Africa: An Emerging Force for Autocrats and Activists," in *Media and Mass Atrocity: The Rwanda Genocide and Beyond*, ed. Allan Thompson (Waterloo, ON: Center for International Governance Innovation, 2019).

[87] Luke Munn, "Angry by Design: Toxic Communication and Technical Architectures," *Humanities and Social Sciences Communications* 7, no. 1 (2020): 1-11; Steve Rathje, Jay J. Van Bavel, and Sander van der Linden, "Out-Group Animosity Drives Engagement on Social Media," Psychological and Cognitive Sciences 118, no. 26 (2020); William Brady et al., "How Social Learning Amplifies Moral Outrage Expression in Online Social Networks," *Science Advances* 7, no. 33 (2021); Andrea Bellovary, Nathaniel Youn, and Amit Goldenberg, "Left- and Right-Leaning News Organizations Use Negative Emotional Content and Elicit User Engagement Similarly," *Affective Science* 2 (2021): 391–396; Jeremy B. Merrill and Will Oremus, "Five Points for Anger, One for a 'Like': How Facebook's Formula Fostered Rage and Misinformation," *Washington Post* (October 26, 2021), https://www.washingtonpost.com/technology/2021/10/26/facebook-angry-emoji-algorithm/;
Matthew Shaer, "What Emotion Goes Viral the Fastest?," *Smithsonian Magazine* (April 2014), https://www.smithsonianmag.com/science-nature/what-emotion-goes-viral-fastest-180950182/; Keith Zubrow, "Facebook Whistleblower Says Company Incentivizes 'Angry, Polarizing, Divisive Content," *CBS News* (October 4, 2021), https://www.cbsnews.com/news/facebook-whistleblower-frances-haugen-60-minutes-polarizing-divisive-content/.

"Social media companies have evolved rapidly and continuously, and expanded quickly into many different countries, disrupting infosystems in ways that can be hard for regulators and human rights monitors to address quickly...

At the same time, [these] platforms have become increasingly thorough at collecting user data, which is used to tailor each user's content feed to maximize their engagement."

—

Matthew MacDonald
Technology and Human Rights Fellow

since 2013 in South Sudan.[88] Although misinformation of this kind isn't new, it is increasing in speed and scale as a result of social media; and this phenomenon is the explicable result of platforms' core incentives—it is not an aberration, but a natural consequence of how the technologies and business models operate. Accompanied by advances in AI-generated images and text, the trajectory is towards more misinformation, not less.

This motivates the thought that some important ethical rights might need to be re-examined. In the first part of this paper I offered a conception of rights that I think best explains their most important features, such as their universality, enforceability, and resistance to tradeoffs. Nonetheless that conception places rights at an instrumental level of moral theorizing, taking them to result from calculations about which entitlements are necessary to protect our basic interests against systematic threats. Because the systematicity of threats can change over time, rights can require occasional recalibration, and new rights can even emerge. The observation that inflammatory misinformation is on the rise, becoming increasingly systematic and contributing to substantial harm, should make us wonder whether our rights as currently formulated are properly drawn.

"The observation that inflammatory misinformation is on the rise, becoming increasingly systematic and contributing to substantial harm, should make us wonder whether our rights as currently formulated are properly drawn."

Although online speech environments bear on multiple rights, chief among them is the right to freedom of speech.[89] This right has a long history and substantial legal and philosophical commentary, but it is generally thought to protect both speaker-side and hearer-side interests. Persons have interests in being able to discuss ideas, refine their opinions, and express themselves autonomously. Hearers have interests in receiving information from a diverse variety of sources, and in being able to think autonomously for themselves, rather than being forced by authorities to accept a predetermined view. Yet both speaker-autonomy, hearer-autonomy, and the safety and autonomy of third parties can be affected by sufficiently inflammatory speech. One of the most extreme ways to remove a person's autonomy is to kill them, so appeals to autonomy will be insufficient to justify extending free speech protections to those forms of speech that systematically cause death. This is similar to the already-accepted rationale for well-known exceptions to free speech, such as dangerously false advertising, dangerously false product labeling, direct incitement to violence, "fighting words," some forms of libel, and fraud. Atrocity-misinformation of the kind described above does not differ in relevant ways from these speech-types, except for the fact that it is not as widely recognized as an exemption from category of protected speech.

It is also arguable that one's autonomy is undermined when one is manipulated into believing misinformation, and that intellectual autonomy (the "capacity to think for oneself") depends on a variety of limited cognitive resources, such as time, expertise, access to evidence, attention, and so forth. The late justice Antonin Scalia, dissenting in *McConnell v. Federal Election Commission* (2003), argued that "[t]he premise of the First Amendment is that the American people are neither sheep nor fools, and hence fully capable of considering both the substance of the speech presented to them and its proximate and ultimate source." Apart from the fact that this is an empirical claim, which is easier to support in some contexts than others, my view is that properly respecting a person's capacity to do something requires recognizing ways in which the capacity may be limited, or its exercise unjustly burdensome. One does not adequately respect hearers' intellectual autonomy when one saturates them with time-consuming, hateful, fear-inducing misinformation and then tells them to think for themselves.

These are some of the considerations that, I think, motivate the thought that adjustments may be needed to the right to free speech. They are not supposed to be a decisive argument; and, in any case, such an argument would have to engage more closely with the substantial legal and philosophical literatures on free speech than I have space to do here.[90] Rather, they are reasons why I think an inquiry of that type is necessary. As I have said, the calculation (or recalculation) of a right's content can be a substantial task, for it must deliver a "close-to-all-things-considered" long-term judgement about the systematicity of the relevant threats and both the necessity and tolerability of the relevant entitlements. What I hope to have given are reasons for thinking

[88] MacLellan, "Fake News, Dangerous Speech and Mass Violence."

[89] Other rights include privacy, rights to self-government (in cases of election interference), anti-defamation rights, etc.

[90] See my forthcoming doctoral dissertation for a discussion of that kind.

that such a calculation is possible (given the conception of rights I have offered), and worth doing.

What adjustments to the free speech right do I have in mind? My proposal is twofold. First, I think it is possible to delineate a particular category of misinformation, whose content can be demonstrated sufficiently clearly to be untrue (or, at least, unsupported by credible evidence), and sufficiently likely to cause grievous harm to a particular target. These two features already characterize the free speech exceptions I mentioned (dangerously false advertising, libel, and fraud)—the thought is simply that the category should be widened slightly to include other forms of widespread misinformation with the same features, such as RTLM's accusations of Tutsi cannibalism, false accusations of widespread Rohingya terrorism, or the various fake reports that appeared on Serbian television during the 1990's, alleging that Bosnians and Croatians were feeding babies to zoo animals.

"In the early phases of the Rwandan Genocide, local UN peacekeeping commander Romeo Dallaire fervently requested US equipment with which to jam the RTLM transmissions: the response from the State Department was a refusal on the grounds that doing so would violate free speech... If harmful misinformation is not in fact covered by such a right, at the very least that fact must become more widely recognized."

One might think that these cases are *already* clearly excluded: that whatever else the free speech right protects, surely it already exempts widespread misinformation about atrocities. But this is not universally agreed upon. In the early phases of the Rwandan Genocide, local UN peacekeeping commander Romeo Dallaire fervently requested US equipment with which to jam the RTLM transmissions: the response from the State Department was a refusal on the grounds that doing so would violate free speech.[91, 92] This deadly decision continues to be repeated in the policies of social media companies, which have been notoriously hesitant about moderating harmful misinformation (especially in non-English languages), and which typically justify these decisions by appeal to the (ethical) free speech right.[93] If harmful misinformation is not in fact covered by such a right, at the very least that fact must become more widely recognized. Explicitly recalculating the right, in ways that better recognize the increased systematicity of hateful misinformation, would help to show why exactly this type of speech is not automatically protected.

My second proposal is more radical. At present, speech categories like false advertising, libel, and fraud are understood to be disunited exceptions to the category of rights-protected speech. In addition to narrowing that category to exclude the most egregious forms of atrocity-misinformation, my proposal is that the exceptions themselves should be considered a unified speech-category, and that we should seriously investigate whether this category can be targeted by its own distinctive right. This would be an anti-misinformation right: a right which defines in more detail what the *epistemic* qualities of these speech types are (that is, specifically which evidential standards they fall short of), and the nature of the harm that they cause (that is, which basic interests they affect). It would also interlock with the (reformulated) right to free speech. That is, the two rights would jointly explain which kinds of speech are and are not open for governments, corporations, and other authorities to regulate, and upon what justifications they may do so. They would also pick out those kinds of speech that authorities are obligated to regulate given what is known empirically about their harmfulness.

[91] Roméo Dallaire, "The Media and the Rwanda Genocide," in *Media and Mass Atrocity: The Rwanda Genocide and Beyond*, ed. Allan Thompson (Waterloo, ON: Center for International Governance Innovation, 2019).

[92] Yanagizawa-Drott), "Propaganda and Conflict."

[93] Sean Illing, "The First Amendment Has a Facebook Problem," *Vox* (May 5, 2021), https://www.vox.com/policy-and-politics/22356339/free-speech-facebook-twitter-big-tech-first-amendment; *Associated Press*, "Zuckerberg Says Facebook Must Stand Up for Free Speech" (February 1, 2020), https://apnews.com/general-news-c3291531831d19ff0eaf8d91aa1415a0; Siva Vaidhyanathan, "Mark Zuckerberg Doesn't Understand Free Speech in the 21st Century," The Guardian (October 18, 2019), https://www.theguardian.com/commentisfree/2019/oct/18/mark-zuckerberg-free-speech-21st-century; Andrew Marantz, "Facebook and the 'Free Speech' Excuse," *The New Yorker* (October 31, 2019), https://www.newyorker.com/news/daily-comment/facebook-and-the-free-speech-excuse.

8. Conclusion

Both of these proposals are tentative, and I have developed neither of them in sufficient detail here. What I hope to have given is a general conception of ethical rights that explains their most important features, and which is useful for those who examine the ethics of new technologies through the lens of human rights. This conception entails that rights can evolve over time, and that new rights can emerge in response to new systematic threats, or changes in the systematicity of existing threats. I gave some historical cases to illustrate that, as a general matter, technological changes can drastically affect infosystems, and the mass communication of hateful misinformation can lead to violence, at least in certain circumstances. My view is that social media platforms contribute to making hateful misinformation more systematic by enabling it to spread at greater speed and scale, and in forms that are harder for users to spot. The problem is worst in languages other than English, and the violent effects are most prevalent in the Global South.

> "...Social media platforms contribute to making hateful misinformation more systematic by enabling it to spread at greater speed and scale, and in forms that are harder for users to spot. The problem is worst in languages other than English, and the violent effects are most prevalent in the Global South."

Rural communications tower. Image credit Jay Heike

Because these changes are the natural consequence of the platforms' core technologies and business models, they are deeply rooted and unlikely to change easily. Combined with developments in AI, the trajectory is towards more misinformation, not less. Consequently, the empirical components in the calculation of a free speech right need updating, in ways that recognize the increased potential of inflammatory speech to harm the interests that rights are designed to protect. These interests include the preservation of life and physical safety, and the autonomy-interests of speakers and hearers. Although I have not attempted to perform the full recalculation of a free speech right in this paper, I have described what I think it would yield: clarifying and limiting the category of protected speech, in ways that recognize the similarity between hateful misinformation and speech types that are already excluded from protection. My view is that we should explore seriously the possibility that these speech types themselves form a unified category, with distinctive epistemic and empirical features, and the possibility that they can be targeted by a distinctive right not to be misled.[94]

[94] I take up that project in my forthcoming dissertation.

Carr Center for Human Rights Policy
Harvard Kennedy School
79 JFK Street
Cambridge, MA 02138

Statements and views expressed in this report are solely those of the author and do not imply endorsement by Harvard University, the Harvard Kennedy School, or the Carr Center for Human Rights Policy.

This publication was published by the Carr Center for Human Rights Policy at the John F. Kennedy School of Government at Harvard University

carrcenter.hks.harvard.edu
79 JFK Street | Cambridge, MA 02138
carrcenter@hks.harvard.edu