Fostering the Intellectual Virtue of Civility in Online Contexts

Deborah S. Mower*

Civility provides norms for engaging in discourse on topics that yield harm or benefit to others, affect the conditions of our lives, and drive public policy—in short, on topics that matter, morally and politically. While discourse on moral and political matters is already difficult in face-to-face and interpersonal conversations, online interactions through social media pose unique challenges for interactions and exchanges. Using research from philosophy, psychology, behavioral economics, and computer science, I explain how the intellectual virtue of civility can be fostered within individuals as a developed trait of internalized norms to guide how one engages with others when evaluating beliefs and contentious issues. Like the development of any virtue, the development of civility takes time, practice, and effort, but it can be fostered even in online contexts through the careful contextualization of experience and practiced behaviors. Altering tech design through the use of goal setting, nudges, gamification, and in-group priming allows for repeatable behaviors to be practiced and modified. Shifting from a focus on policies for content moderation to a focus on users' choices within online interactions—and deeper consideration for how designs affect users' options, choices, and agency—allows us to capitalize on insights from moral psychology and education to see new ways forward in combatting incivility in online contexts.

INTRODUCTION

Disagreement on issues is currently so widespread that there is perhaps only one thing upon which we agree: online contexts in general—and social media in particular—have hurt our social interactions and the health of our democracy.[1] The challenges of online contexts are well known. Throughout

* Associate Professor of Ethics endowed by Mr. and Mrs. Alfred Hume Bryant, Department of Philosophy and Religion, University of Mississippi.

1. An earlier version of this paper was presented at the Tech Law and the Humanities Symposium at the Yale Law School in November 2022. I am grateful to members of the audience for their thoughtful questions and insights as well as the support of The Justice Collaboratory of the Yale Law School. Special thanks to Paul Meosky for his extraordinary work in coordinating the symposium and to the editors of the *Yale Journal of Law & the Humanities*, and Jiawei Wang, for the helpful comments

most of history, one's social groups were largely circumscribed by in-person associations: regional, political, familial, religious, educational, and civic.[2] But the rise of social media has allowed our associations and networks to have a uniquely unprescribed character. The highly voluntary and fluid nature of our online social networks yields limited accountability as individuals can mute comments from others, break contact with or remove others from their networks, and gain access to new networks—or leave them—with the click of a button. As is well documented by now, we sort ourselves into like-minded groups where much of what we say and profess to believe is shared by our audience. We can limit our exposure to views with which we disagree, individuals that we find to be morally reprehensible, and interactions that we do not enjoy, while seeking unanimity with others.[3] Online contexts and social media have changed our social experiences and how it is possible to engage—or fail to engage—with others. Further, individuals can share and post content with complete anonymity, letting loose their inner demons and spreading venom without checks. We have collectively created the conditions that render accountability a fiction. And because of the voluntary and fluid nature of our associations online, there is no guarantee—or even expectation—for repeated exchanges or interactions over time. The flux of our social experience online fractures former bonds of cohesion and trust developed through practiced interpretation, understanding, and compromise within stable group contexts.[4] While these particular challenges of online contexts lead to a variety of social ills such as misinformation, cyberbullying, polarization, incivility, and silencing, my focus in this paper is on incivility as a unique problem—a problem that either directly causes or causally

throughout the editing process.

2. MacIntyre, Alasdair. After Virtue, 3rd Edition. Notre Dame, Indiana: University of Notre Dame Press, 2007.

3. Lord, Charles, Ross, Lee, and Mark Lepper. "Biased Assimilation and Attitude Polarization: The Effects of Prior Theories on Subsequently Considered Evidence." *Journal of Personality and Social Psychology* 37 (1979): 2098–2109; Iyengar, Shanto, Gaurav Sood, and Yphtach Lelkes. "Affect, Not Ideology: A Social Identity Perspective on Polarization," *Public Opinion Quarterly* 76 (2012): 405-431; Stroud, Natalie. "Polarization and Partisan Selective Exposure." *Journal of Communication* 60 (2010): 556-576.

4. Mason, Lilliana. "'I Disrespectfully Agree': The Differential Effects of Partisan Sorting on Social and Issue Polarization." *American Journal of Political Science* 59 (2015): 128-145; Törnberg, Petter, "How Digital Media Drive Affective Polarization Through Partisan Sorting." *Proceedings of the National Academy of Sciences (PNAS)* 119 (2022): 1-11; Anheier, Helmut and Jeremy Kendall. "Interpersonal Trust and Voluntary Associations: Examining Three Approaches." *British Journal of Sociology* 53 (2002): 343-362. Kwon, K. Hazel, Chun Shao, and Seungahn Nah. "Localized Social Media and Civic Life: Motivations, Trust, and Civic Participation in Local Community Contexts." *Journal of Information Technology & Politics* 18 (2021): 55-69. The fragmentation of historic associational networks and the erosion of interpersonal trust can be counteracted (to some extent) through the creation of localized social media (LSM) networks (such as the neighborhood website/social app Nextdoor), demonstrating the importance of intentional purpose and design for maintaining stable social contexts.

contributes to each of these other social ills—and ways to foster civility as a solution.

Prevailing approaches to the problem of incivility in online contexts have treated it as either a legal or an institutional/organizational issue.[5] In treating it as a legal issue, the method is to examine the content and determine whether it falls within the legally protected category of free speech or the morally condemnable category of hate speech. This exercise is one of examining legal rulings and statutes, evaluating various philosophical and legal arguments about what qualifies as hate speech, interpreting the content, and rendering a judgment. In treating it as an institutional/organizational issue, the method is to examine the content and determine whether it follows or runs afoul of internal policies. But of course, given the rapid growth and changes in online contexts, institutions often find that they have no relevant policy for guidance. This exercise is one of examining or creating internal policy, evaluating the dynamics, goals, and services of the institution, interpreting the content in light of that institutional framing, and rendering a judgment. The task assumed by both approaches is how to guide or moderate content via policy formation and application (whether as law or internal rules). The outcome has been a cottage industry of scholarship on hate speech and free speech in online contexts,[6] and an avalanche of community standards, behavioral standards, and guidelines for content issued by various social media companies.[7] Individual users are responsible for knowing and following the policies, and failure to do so carries punishments such as content removal or account

5. One notable exception is research by Seering et al. on how to redesign the graphic user interface to prime positive emotion using CAPTCHA images, increasing analytical complexity and social connectedness. Seering, Joseph, Tianmi Fang, Luca Damasco, Mianhong Chen, Likang Sun, Geoff Kaufman. "Designing User Interface Elements to Improve the Quality and Civility of Discourse in Online Commenting Behaviors." *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing System* (New York, NY: Association for Computing Machinery, 2019), 1-14.

6. Gagliardone, Iginio, Danit Gal, Thiago Alves, and Gabriela Martinez. *Countering Online Hate Speech*. France, UNESCO Publishing, 2015; Tsesis, Alexander. "Hate in Cyberspace: Regulating Hate Speech on the Internet." *San Diego Law Review* 38 (2001): 817-874; Whitman, James Q. "Enforcing Civility and Respect: Three Societies." *Yale Law Journal* 109 (2000): 1279-1398; Council of Europe, "Additional Protocol to the Convention on Cybercrime, Concerning the Criminalisation of Acts of a Racist and Xenophobic Nature Committed Through Computer Systems." *European Treaty Series* 189 (2003): 1-6. For more information on the protocol issued by the Council of Europe (including countries that have signed the treaty), visit: https://www.coe.int/en/web/conventions/full-list?module=treaty-detail&treatynum=189 .

7. For example, see the 2016 "Code of Conduct on Countering Illegal Hate Speech Online" agreement between the European Commission and four IT companies (Facebook Microsoft, Twitter, and YouTube) at https://commission.europa.eu/document/551c44da-baae-4692-9e7d-52d20c04e0e2_en . The IT companies agreed to develop internal policies for content moderation, educate and notify users of rules and community guidelines, and to institute processes for review, removal of, or restricted access to illegal hate speech within a 24 hour period. See YouTube's Community Guidelines at https://www.youtube.com/howyoutubeworks/policies/community-guidelines/ . See Facebook's Community Standards at https://transparency.fb.com/policies/community-standards/?source=https%3A%2F%2Fwww.facebook.com%2Fcommunitystandards#hate-speech. See Twitter's Terms of Service at https://twitter.com/tos?lang=en#usContent .

suspension. Given the widespread increase in incivility in the last decade, it is manifestly clear that these approaches have not been particularly effective.

The ineffectiveness of these approaches for exchanges in online contexts stems from their broadly exclusionary and abstract nature. Many rules, policies, and laws offer clear limits on actions or behaviors. For example, speed limits set a maximum rate of travel, such as a posted speed of 55 mph. Any speed above that boundary number is prohibited and individuals traveling at higher rates of speed are subject to a ticket if caught. The posting excludes speeds above a boundary limit but does not offer guidance to drivers on more appropriate slower speeds for driver distraction, road conditions, lighting, or traffic levels. Similarly, former approaches to the problem of incivility are broadly exclusionary. While they seek to limit hate speech or that which excites violence, they are silent and offer no guidance for exchanges that fall below that exclusionary and boundary limit. Merely promulgating limiting policies on civility, content, and speech provides neither guidance nor opportunity for meaningful agency. Presenting users with a collection of rules, policies, and laws presents them with choice options within a limited range, but what is needed is a deeper consideration of the conditions of those options for choice and behavioral outcomes more broadly.

Rather than thinking of incivility in online contexts as a legal or institutional/organizational problem, we need to reconceive it as a problem at the level of the individual—more specifically, at the level of individual psychology, in how individuals make choices for their interactions with others. Civility provides norms for engaging in discourse on topics that yield harm or benefit to others, affect the conditions of our lives, and drive public policy—in short, on topics that *matter*, morally and politically. Although social media poses unique challenges for interactions and the exchange of ideas, I argue that the intellectual virtue of civility can be fostered within individuals as a developed trait of internalized norms even within online contexts. This shift to the psychological allows us to focus on how situations structure choices and provide opportunities to inform and educate users. Most importantly, the focus on civility as an intellectual virtue allows us to consider how technology policies and designs can promote civil discourse between citizens and foster civility within individuals—or fail to do so and breed discord and incivility.

In this paper, I first explain the intellectual virtue of civility. Using several examples, I explain how civility norms are internalized through goal selection and action, opportunities for direction and correction, and then instilled into habit through repetitive practice. I then explain how virtues are cultivated over time through the careful construction of situations to help agents make choices and develop positive intellectual traits. Using research

from philosophy, psychology, behavioral economics, and computer science, I discuss how both non-conscious and conscious processing play important roles in our choices and behavior. Priming, goal automaticity, nudges, and gamification collectively serve to inculcate virtue through motivating action, providing subtle yet powerful direction and correction, and repeated opportunities for practice and learning. In the final section, I sketch multiple proposals for how online platforms could implement these tools, showing how automaticity, priming, nudges, and gamification each have educational roles in shaping users' developing skills for civility. Civility, like all intellectual virtues, can be developed over time through careful contextualization of experience and practiced behaviors, and it can be fostered even in online contexts. Shifting from a focus on policies for content moderation to redesigning how users craft content, post, and interact with others provides structured opportunities to improve as well as concrete guidance for civility.

THE INTELLECTUAL VIRTUE OF CIVILITY

There are a number of educational resources that we can appeal to from literature within philosophy and moral psychology that can help us understand the importance of this shift. Ethicists have done extensive work on the nature of virtue,[8] how it can be developed within educational settings,[9] what it means for virtues to be developed or instilled within thinkers and learners,[10] and how developed virtues guide moral action.[11] We can begin by considering an example of how assumed and internalized shared norms govern a social or group activity and how they become internalized through social feedback and practice.

Think of a game of touch football, which is, of course, a common family activity after holidays such as Thanksgiving dinner. For my family to be able to play, we must have some shared norms or points of agreement. Without these following agreements (which are often mere assumptions and never explicitly stated), we would not be able to engage or participate in the activity at all. First, we must agree that there are two teams (and not four,

8. MacIntyre, Alasdair. *After Virtue*. Notre Dame, IN: University of Notre Dame Press, 1981; Hursthouse, Rosalind. *On Virtue Ethics.* New York, NY: Oxford University Press, 1999; Slote, Michael. *From Morality to Virtue*. New York, NY: Oxford University Press, 1992; Battaly, Heather. *Virtue.* Malden, MA: Polity, 2015.

9. Baehr, Jason. *Deep in Thought: A Practical Guide to Teaching for Intellectual Virtues*. Cambridge, MA: Harvard Education Press, 2021; Curren, Randall. "Virtue Epistemology and Education." In *The Routledge Handbook of Virtue Epistemology*, ed. Heather Battaly (New York, NY: Routledge, 2020): 470-482;

10. Zagzebski, Linda. *Virtues of the Mind: An Inquiry into the Nature of Virtue and the Ethical Foundations of Knowledge.* Cambridge, UK: Cambridge University Press, 1996; Wilson, Alan and Christian Miller. "Virtue Epistemology and Developing Intellectual Virtue." In *The Routledge Handbook of Virtue Epistemology*, ed. Heather Battaly (New York, NY: Routledge, 2020): 483-495;

11. Annas, Julia. *Intelligent Virtue*. New York, NY: Oxford University Press, 2011.

six, or each person as an individual team of one). Clearly, the game cannot proceed if we have multiple competitions at the same time, are playing our own versions of the game, and are not coordinating our actions into a competition. Second, we must agree that there are important and valued locations represented by the end zones. These shared places of value motivate and direct specific actions, such as sprints to the far end of a grassy span or defending areas marked by lines or cones. Third, we must have a standard method for achieving points and tracking scores. We agree that points are gained by running the football across the line, but not by the height achieved while throwing. (Imagine the chaos if one person continually captured the ball and persisted in repeatedly throwing it skyward.) Fourth, we must agree that there are acceptable physical actions. Because this is a game of touch football, not tackle football, actions that violate the accepted parameters are disallowed. For example, if my brother-in-law tackled me, then the game would end because he engaged in an unacceptable physical action. Fifth, we recognize that there are acceptable verbal actions. In the context of touch football, teasing and good-natured ribbing are part of what makes the game enjoyable, but we all recognize that threats are not appropriate. For example, it would be entirely out of bounds for my brother-in-law to threaten to disown me from the family if I touched the football. Such a comment would indicate that he and I are engaging in different activities with different norms. Again, the game would end. And sixth, we recognize that there is a conclusion or an endpoint for the nature of the activity itself. So if my brother-in-law takes the football and runs it across the endzone next week, he does not thereby gain a point or change the outcome of the game from the previous week. These norms collectively guide our interaction so that we can have a competition. More specifically, they provide the format, the understanding of procedure, values and goals, meanings (to make inferences and interpret the behavior of others), and parameters or limits on actions. Without such shared norms, we have no game; we are left with uncoordinated individual action, chaotic and ineffective units, or violent pileups.

With this example in place of shared norms which enable complex group activity, we can illustrate how civility similarly provides internalized and shared norms that govern debate and disagreement.[12] To engage in disagreement or debate over moral and political matters, we also must have shared norms or points of agreement. Without the following agreements (which are also often assumptions and never explicitly stated), we will not

12. Mower, Deborah. "Civility." In *The International Encyclopedia of Ethics, 3rd Edition,* ed. Hugh LaFollette (Hoboken, NJ: Wiley Blackwell, 2021), 1-8; Mower, Deborah. "The Real Morality of Public Discourse: Civility as an Orienting Attitude." In *A Crisis of Civility?: Political Discourse and its Discontents,* eds. Robert Boatright, Sarah Sobieraj, Dannagal Goldthwaite Young, and Timothy J. Shaffer (New York, NY: Routledge, 2019), 210-232.

be able to engage. First, we must recognize that we are evaluating our personal beliefs about the public good, not merely evaluating personal beliefs about our personal actions. Because beliefs guide the actions of others in questions of morality and policy, this is a shared matter just in the same way as a game is a cooperative, shared matter. Second, we agree on the value and method of considering all evidence and sides of an issue, not just a portion of evidence or that side with which I agree. Even individuals who engage in incomplete or motivated reasoning (mistakenly) believe that they have considered all the evidence or engaged in non-biased reasoning.[13] Our success or failure in living up to our values and methods in practice does not lessen our desire for and belief in a gold standard for reasoning methods (in discussions, debates, and law). Third, we share values of consistency and impartiality. We recognize that for a discussion or debate to occur, each party must be given time to contribute, that participants should all have equal opportunity to present their positions, and that views should not be excluded because of who the presenter is or the position for which they stand. Without joint participation, we have a monologue or a platform rather than a discussion or debate. Fourth, similarly to the game of touch football, we have shared motivations and goals. We expect that one states claims fairly and presents evidence fully and honestly because the aim is to uncover the truth.[14] Lastly, we agree that there are behavioral parameters or unacceptable actions such that if they are performed, the conversation, discussion, or debate ceases—the result is an end to the activity. Unacceptable actions include provoking or maligning others, being intolerant, or refusing to engage.

These norms not only guide how we evaluate beliefs, but also how to interact with each other when doing so. They collectively provide the format, understanding of procedure, values and goals, meanings (allowing for inference and ways to interpret the behavior of others), and parameters/limits on actions. They guide our interactions so that we are able to have an actual discussion and debate, rather than shouting at each other, talking past each other, or simply refusing to engage. Just as the internalized norms for the shared group activity of touch football enable me to play with my family (or anyone in novel circumstances), civility enables one to engage in both private and public discourse to evaluate belief—especially the highly complex and fraught beliefs we care most about: moral and

13. Baumeister, Roy and Leonard Newman. "Self-regulation of Cognitive Inference and Decision Processes." *Personality and Social Psychology Bulletin* 20 (1994): 3-19; Ehrlinger, Joyce and David Dunning. "How Chronic Self-views Influence (and Potentially Mislead) Estimates of Performance" *Journal of Personality and Social Psychology*, 84 (2003): 5- 17. Kunda, Ziva. "The Case for Motivated Reasoning." *Psychological Bulletin* 108 (1990): 480-498.

14. Levine, Timothy, Narissa Punyanunt-Carter, and Alivia Moore. "The Truth-Default and Video Clips: Testing the Limits of Credulity." *Communication Studies* 72 (2020): 133-145; Levine, Timothy. "Truth-Default Theory (TDT): A Theory of Human Deception and Deception Detection." *Journal of Language and Social Psychology* 33 (2014): 1-15.

political matters.

While norms are often assumed and left unstated precisely because they have already been internalized, virtue theorists focus on the process of how virtues are developed through activities involving norms, feedback through direct instruction or correction, and then habituated through practice.[15] To understand more about how virtues develop into traits, consider the goal of learning to play the piano. Learning to play involves engaging in many activities and learning new actions. One must learn what various symbols and bars on sheet music mean, how to decipher chords, and interpret marks for tempo and volume. The next step is learning to pair the written music symbols with the patterns of black and white keys on the board, followed by the repeated process of translating those symbols to keys through finger placement, movement, rhythm, and timing. One internalizes the meaning and guidance of written music through learning to read the notes and translating that knowledge into action on the keyboard. Being able to hear the semblance of a tune emanating from one's fingers yields a remarkable sense of accomplishment. This success spurs more playing, more time, and more learning; success both drives and feeds intrinsic motivation. But being able to play is not being able to play *well*. Piano lessons provide the opportunity for improvement and the achievement of excellence through focused feedback. A piano teacher provides specific instructions, recommendations for alteration, and feedback—both positive and negative—on one's performance. Hearing other players perform at recitals and concerts provides models of excellence, and a teacher's constant direction and correction through weekly practices provide external motivation for further improvement. With sufficient practice and instruction, one no longer pauses and fumbles notes, but plays well, having faithfully mastered transforming written sheets of notes into music. And with repeated practice over time, one begins to play with effortless action, greater creativity, and improvisation. The process of developing virtue is similar to that of learning to play the piano: engaging in and absorbing the norms of an activity, developing component and concrete skills through instruction, and cultivating developed traits of expertise.

Educational contexts provide clear examples and models for developing intellectual virtues. Intellectual virtues are those qualities of mind and developed character needed to think critically, seek understanding, and pursue truth—in short, to be a good thinker and learner. Curiosity, open-mindedness, intellectual humility, tenacity, intellectual courage, and civility enable individuals to seek out more information to challenge their beliefs, engage with others to help clarify and expand their knowledge and learning,

15. Stichter, Matt. "Virtues, Skills, and Right Action." *Ethical Theory and Moral Practice* 14: 73-86; Dreyfus, Hubert and Stuart Dreyfus. "Toward a Phenomenology of Ethical Expertise." *Human Studies* 14: 229-250.

and to increase their intellectual growth and excellence.[16] These intellectual virtues not only support becoming better thinkers, and thereby better citizens, but also aid how we make moral choices and navigate complex political dynamics and social interactions to become more moral citizens.[17]

While we cultivate civility throughout the undergraduate and graduate curricula, humanities courses provide the most direct instruction and opportunities for practice given their historical, moral, religious, and political content.[18] We teach students the norms of civility for how to engage in the learning, discussion, and study of contentious topics. For example, we support the goal of seeking knowledge and teach the process of belief evaluation in filtering and evaluating evidence in support of the reasons. We teach students the format of verbal exchange and questioning through regular class conversations, and the method of considering all sides of issues by having students present varied perspectives within classroom conversations and insisting on multiple sources of evidence in their writing projects. We show students how to conduct quality research, how to use their findings to back up the claims that they make, and how to challenge other research. We also teach them that personal attacks on speakers or authors, fallacious arguments, and a disregard for academic and scientific knowledge are unacceptable. We provide direction through specific instructions and corrective guidance on both their oral and written assignments, and hone their skills through the processes of presentations, writing, and revision. These norms of civility become habituated through such repeated practice and continual feedback, stretching across thousands of discrete experiences in years of class sessions, moments of study and research, and conversations with peers and professors. Just as with learning to play the piano, some individuals barely achieve rudimentary abilities, others are able to engage in discussions and debates on contentious issues, while still others become quite skilled.

While students gain robust models for civility as well as extensive support

16. Baehr, Jason. *The Inquiring Mind: On Intellectual Virtues & Virtue Epistemology*. New York, NY: Oxford University Press, 2011; Kidd, Ian. "Educating for Intellectual Humility." In *Intellectual Virtues and Education: Essays in Applied Virtue Epistemology*, Jason Baehr (ed.) (New York, NY: Routledge, 2017), 54-70; Riggs, Wayne. "Open-Mindedness, Insight, and Understanding." In *Intellectual Virtues and Education: Essays in Applied Virtue Epistemology*, Jason Baehr (ed.) (New York, NY: Routledge, 2017), 18-37.

17. King, Nathan. *The Excellent Mind: Intellectual Virtues for Everyday Life*. New York, NY: Oxford University Press, 2021; Frierson, Patrick. "Intellectual Virtues," In *Intellectual Agency and Virtue Epistemology: A Montessori Perspective* (New York, NY: Bloomsbury Academic, 2021), 59-87; Roberts, Robert and W. Jay Wood. *Intellectual Virtues: An Essay in Regulative Epistemology*. New York, NY: Oxford University Press, 2007; Garcia, Robert and Nathan King. "Toward Intellectually Virtuous Discourse." In *Intellectual Virtues and Education: Essays in Applied Virtue Epistemology*, Jason Baehr (ed.) (New York, NY: Routledge, 2017), 202-220.

18. Mower, Deborah. "Moderating Conviction through Civility in Education." In *Bias, Belief, and Conviction in an Age of Fake Facts*, eds. Anke Finger, and Manuela Wagner (New York, NY: Routledge, 2021), 134-155; Mower, Deborah and Wade Robison (eds.). *Civility in Politics and Education*. New York, NY: Routledge, 2012.

and development throughout their educational experience, their capacity is either reinforced or eroded within their broader social relations and interactions—whether in-person with parents and peers or online with associated friends and strangers. Because social contexts provide myriad opportunities to shape and train actions and choices, the same process of learning and engaging in an activity, receiving direction and correction, and habituation through practice can cultivate either virtue or vice. To understand how the same developmental process can result in vice, consider the example of lying. One easily picks up forms and methods of lying, and can become a good liar if given opportunities to practice. Negative influences (such as peers) or inducements to lie (such as the possibility of being punished for truth-telling) provide direction and motivation. And further direction and correction often occur through positive reinforcement of negative behavior when a liar benefits from successful deception. Over time, such ingrained action becomes easy to perform. We all know examples of individuals who are consummate liars—accomplished virtuosos—who are so skilled that they not only delude their listeners but also themselves. The fact that virtue and vice are developed via the same process of learning, feedback, and habituation highlights the importance of attending carefully to the contexts and situations that either cultivate or erode civility.

CULTIVATING VIRTUE THROUGH SITUATIONAL CONSTRUCTION AND CHOICE

With this understanding of the intellectual virtue of civility in place, now we can examine the construction of situations and choices more closely. Learning and development of virtue occur through both conscious and unconscious processes, and we can cultivate virtue through both processes. Virtue theorists have long known that the careful construction of situational features helps to train individuals to respond appropriately—affectively, cognitively, and behaviorally—even when the situational features are below the level of conscious attention.[19] To see why situational features are so powerful, we next turn to some of the extensive psychological research on non-conscious processing.

19. Brady, Michael. "The Role of Emotion in Intellectual Virtue." In *The Routledge Handbook of Virtue Epistemology*, ed. Heather Battaly (New York, NY: Routledge, 2020): 47-57; Hutton, Erik. "Character, Situationism, and Early Confucian Thought." *Philosophical Studies* 127 (2006): 37-58; Frierson, Patrick. "Unconscious and Embodied Intellectual Agency," In *Intellectual Agency and Virtue Epistemology: A Montessori Perspective* (New York, NY: Bloomsbury Academic, 2021), 35-58; Merritt, Maria. "Virtue Ethics and Situationist Personality Psychology." *Ethical Theory and Moral Practice* 3 (2000): 365-383; Mower, Deborah. "Situationism and Confucian Virtue Ethics." *Ethical Theory and Moral Practice* 16 (2013): 113-137; Mower, Deborah. "Scripting Situations in Moral Education." *Teaching Ethics* 11 (2010): 93-106; Snow, Nancy. *Virtue as Social Intelligence: An Empirically Grounded Theory*. New York, NY: Routledge, 2009; Slingerland, Edward. "The Situationist Critique and Early Confucian Virtue Ethics." *Ethics* 121 (2011): 390-419.

Our ideas are linked in networks of associations, both consciously and unconsciously. Presenting subjects with a specific image, word, or concept will make related ideas more accessible, meaning that they have an increased probability of access, or use, in thought, speech, or action. Imagine for a moment the image of a heart that might be appropriate for a Hallmark card. Given that image, the words "heart," "love," or "valentine" are undoubtedly more likely to come to one's mind than words such as "donor," "organ," or "surgery." But now, imagine a new image: one of a human heart covered in veins and appropriate for a medical textbook. Given this new image, words such as "surgery," or "organ" are highly likely to come to one's mind while "valentine" would be quite surprising. We can capitalize upon this pairing between images and concepts, and the associations between them, by using stimuli to target specific conceptual associations. This process is known as priming, and can be defined as increasing the conceptual accessibility of a target concept through the stimulus of associated words or concepts below the level of conscious awareness. Consequently, one clear way that we can construct situations is to prime ideas or concepts and increase the accessibility of related target ideas.[20]

Although not as well-known as the priming research on images and linguistic associations, there is also extensive work on priming behaviors through complexes of related ideas. John Bargh and his research colleagues[21] provide a nice example of a study on trait construct priming. Like most psychology studies, the experimenters gave subjects a distractor task to help limit their awareness of study goals. In the distractor task, subjects were given a collection of scrambled words and asked to re-order the words into grammatical sentences. The subjects were each randomly placed into three separate conditions: polite, neutral, and rude. Individuals in the polite or respect-based condition received terms and words that are typical synonyms for respect such as "patiently," "courteous," and "considerate." Individuals in the rude condition received terms and words that are typical synonyms for rudeness such as "bother," "bluntly," and "intrude." Individuals in the neutral condition received a variety of terms that are not synonyms for either of the other two concepts, such as "prepares," "occasionally," and "exercising." After completing the scrambled sentence task, subjects were instructed to seek out the

20. Bargh, John. "What Have We Been Priming All These Years? On the Development, Mechanisms, and Ecology of Nonconscious Social Behavior." *European Journal of Social Psychology* 36 (2006): 147-168.

21. Bargh, John, Mark Chen, and Lara Burrows. "Automaticity of Social Behavior: Direct Effects of Trait Construct and Stereotype Activation on Action." *Journal of Personality and Social Psychology* 71 (1996): 230-244; Bargh, John and Paula Pietromonaco. "Automatic Information Processing and Social Perception: The Influence of Trait Information Presented Outside of Conscious Awareness on Impression Formation." *Journal of Personality and Social Psychology* 43 (1982): 437-449.

experimenter in another room to complete the second task of the study. A fellow researcher in the second room pretended to be a confused study participant, asking many questions and misunderstanding directions in a lengthy conversation. The experimenter measured the amount of time it took for each subject to interrupt the conversation. Approximately 65% of subjects in the rude condition interrupted the conversation, compared to below 40% interruption for those in the neutral condition and below 20% for those primed with respect-related words. Research such as this demonstrates that it is possible to prime concept complexes and thereby increase the probability of targeted behaviors.

There is also robust research on the effect of primed group membership on our behavior. Extensive research on group dynamics and social identity theory over the last fifty to sixty years provides robust findings that our behaviors toward others are easily primed by whether we perceive them to have the same identity as ourselves.[22] Researchers refer to the perception of others as being members of the same group as ourselves as an "in-group condition." When we perceive others as being members of the "same" group, we view them as sharing the same characteristics as ourselves. We also evaluate them more favorably on many factors, such as higher levels of attractiveness, trustworthiness, being more intelligent, more justified, better leaders, etc.[23] However, when we perceive others as being members of a "different" group, we view them as having different characteristics than ourselves and evaluate them negatively compared to how we would judge our own qualities or actions (or other members of our group). In-group priming can thus be defined as increasing the salience of perceived similarities between oneself and another through the stimulus of in-group condition (group identity). We can prime particular behaviors by targeting how individuals think of others in relation to their own group membership.

Research from helping behavior studies provides a very nice example of in-group priming. Stürmer and his research colleagues[24] examined the

22. Allport, Gordon. *The Nature of Prejudice*. New York, NY: Basic Books, 1979; Billig, Michael and Henri Tajfel. "Social Categorization and Similarity in Intergroup Behavior." *European Journal of Social Psychology* 3 (1973): 27-52; Brewer, Marilynn. "The Psychology of Prejudice: Ingroup Love or Outgroup Hate?" *Journal of Social Issues* 55 (1999): 429–444; Brewer, Marilyn. "Social Identity, Distinctiveness, and In-Group Homogeneity." *Social Cognition* 11 (1993): 150-164; Brewer, Marilyn. "In-Group Bias in the Minimal Intergroup Situation: A Cognitive-Motivational Analysis." *Psychological Bulletin* 86 (1979): 307-324.

23. Benson, Peter, Stuart Karabenick, and Richard Lerner. "Pretty Pleases: The Effects of Physical Attractiveness, Race, and Sex on Receiving Help." *Journal of Experimental Social Psychology* 12 (1976): 409-415; Williams, Michelle. "In Whom We Trust: Group Membership as an Affective Context for Trust Development." *The Academy of Management Review* 26 (2001): 377-396; Balliet, Daniel, Junhui Wu, and Carsten De Dreu. "Ingroup Favoritism in Cooperation: A Metaanalysis." *Psychological Bulletin* 140 (2014): 1556–1581; Schaller, Mark. "In-group Favoritism and Statistical Reasoning in Social Inference: Implications for Formation and Maintenance of Group Stereotypes." *Journal of Personality and Social Psychology* 63 (1992): 61-74.

24. Stürmer, Stefan, Mark Snyder, Alexandra Kropp, and Birte Siem. "Empathy-Motivated Helping: The Moderating Role of Group Membership." *Personality and Social Psychology Bulletin* 32

impact of group membership on helping behaviors through an innovative investing study. As before, researchers provided subjects with a distractor task to mask the goals of the study. On the basis of their supposed "performance" in this task, researchers informed the subjects that they fell into one of two separate categories: either a detailed or a global perceiver. The researchers then had subjects complete a variety of investing tasks and assigned each a "partner" (a fellow researcher posing as another subject) with whom they would consult throughout the process. The researchers divided the subjects randomly into either an in-group or out-group condition. Those placed in the in-group condition were paired with a partner who shared the same perceiver style (e.g., "detailed" perceiver), whereas those in the out-group condition were matched with a partner from the opposing group. The "partners" introduced themselves to the subjects and mentioned being distracted due to having a bad day (e.g., a lost wallet with money and sporting tickets inside). As one of their investing tasks, subjects selected what to do with their final earnings. One constructed opportunity included the option to donate the money to their partner. After concluding the investing tasks, subjects completed surveys about their process of investing, communication with their partner, and their evaluation of their partner. In-group condition subjects rated their partner as being more similar to themselves, reported greater distress or concern for their plight (e.g., lost wallet and money), and were more likely to help by donating their earnings to their partner than those in the out-group condition. It is clear that we can construct situations to prime in-group stimuli (group identity) and thereby increase more positive evaluations, greater empathy and care, and helpfulness toward a corresponding party. Combined with research on how we view members of our own group with more trust, view them as more intelligent, and afford them a greater benefit of the doubt, there are clear implications for how we can prime civil behavior.

In addition to constructing situations, we also actively construct the conditions of choice. Careful construction of choice options can also inculcate virtue by providing opportunities for the repetition of action into the installation of habit. To see why choice construction is so powerful and how it maintains agency, we turn next to extensive research in psychology and behavioral economics on conscious and rational decision-making.

Choices always have outcomes, and agents select choices and pursue action steps based on their preferences, goals, and values. Suppose that I have the goal of learning to play the piano. Although one does not decide to make a goal automatic, the decision to enact and then pursue a goal can yield automatic behavior or an automatized goal. In the beginning, one must deliberately and consciously decide where to place each finger on a

(2006): 943-956.

keyboard. But over time, the goal of learning to play the piano becomes automatized: one moves fingers smoothly over the keys without ever thinking about implementing the goal or the necessary steps to achieve it. Goals are automatized when the associations of the component attitudes, ideas, affects, mental states, movements, and perceived features of the situation are linked, activated, and strengthened over time.[25] Similarly to priming, automatized goals can be activated by internal or external triggers. For example, if one chooses to engage with civility in situations of disagreement and debate, then the external interactions and tension among the parties and the internal anxiety it yields could trigger the automatized goal of engaging with civility—all without rising to the level of conscious attention, reflection, or overt decision-making. Although the mental processing eventually recedes below conscious attention, the actions result through the automaticity of the previously and consciously chosen goal.

Building on how agents' preferences, goals, and values lead to action, Thaler and Sunstein[26] noted that all situations contain structures of choice and outcome pairings which they refer to as "choice arrays." Agents examine the variety of choices, consider the outcomes, weigh them according to their preferences, goals, and values, and then select a choice that best accords thereby.[27] Returning to the piano example: I can either choose to pay for lessons or try to teach myself with various tools such as books or videos on YouTube. If I have a preference to save money and am unconcerned about time, then I might opt to teach myself. If I have a preference to learn to play quickly and am concerned about maintaining a practice schedule without accountability to another, then I might opt to pay for lessons. Knowing my proclivity to falter with rigorous practice schedules, I recognize that I will need active encouragement from an instructor and so will choose to pay for lessons.

Thaler and Sunstein developed the concept of nudges to capture the idea that within any given situation, agents are inclined or more likely to select some choices over others based on the available choice array. For example, the situational context of visiting my local bank presents a variety of possible choice and outcome pairings regarding deposits, investments, withdrawals, and loans—even thefts. Focusing on the choice arrays within

25. Bargh, John and Erin Williams. "The Automaticity of Social Life." *Current Directions in Psychological Science* 15 (2006): 1-4; Bargh, John, Peter Gollwitzer, Annette Lee-Chai, Kimberly Barndollar, and Trötschel, R Roman. "The Automated Will: Nonconscious Activation and Pursuit of Behavioral Goals." *Journal of Personality and Social Psychology* 81 (2001): 1014-1027; Bargh, John and Melissa Ferguson. "Beyond Behaviorism: On the Automaticity of Higher Mental Processes." *Psychological Bulletin* 126 (2000): 925-945.

26. Thaler, Richard and Cass Sunstein. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. New York, NY: Penguin Books, 2009.

27. Sunstein, Cass. "Nudges Do Not Undermine Human Agency." *Journal of Consumer Policy* 38 (2015): 207-210; Sunstein, Cass; Lucia Reisch, and Micha Kaiser. "Trusting Nudges? Lessons From an International Survey." *Journal of European Public Policy* 26 (2019): 1417-1443.

situational contexts allows us to design outcomes that may appeal more directly to agents and to make the selection of one choice over another easier or more attractive. When a regional bank manager learns that customers are interested in investment management, it would be wise to relocate the office from the back hallway (with advisors available by appointment only) and create a large and visible office for drop-in meetings. In light of customer interest, this simple redesign allows customers to more easily chose options for investment advice. Consequently, nudges are defined as the selection of a choice based on an agent's preferences, goals, or values aligned with an option within a designed choice array that does not incur economic penalties or impose mandates.

While there are numerous studies that now make use of nudges, one of the clearest examples comes from an early study conducted by Thaler and Benartzi.[28] Although there are many types of retirement programs, some are defined contribution plans, in which users voluntarily contribute a percentage of their income toward their employer-managed retirement fund. Savings rates are notoriously low for such plans because individuals have so many other needs for their money: house payments, car repairs, unexpected medical expenses, and hefty grocery bills all seem more pressing than distant retirement needs. Thaler and Benartzi developed the "Save More Tomorrow" program to encourage a higher savings rate for these retirement plan types. They created a program and gave individuals the option to enroll and choose a particular contribution amount. To minimize the hassle of remembering to submit a payment, contributions were automatically deducted from paychecks each month. And to motivate employees to both stay in the program and to voluntarily save more, pay raises were linked to increased contributions. After the program had been in place for 40 months, the average savings rate increase rose from 3.5% to 13.6%—a full 10% increase. The program yielded this remarkable and dramatic improvement by meeting the conditions of a nudge: (1) the contribution was not mandated as an action or in dollar amount, (2) individuals were free to make an alternate decision (they were free to opt out of the program at any point in time), (3) individuals made an explicit choice and commitment to the program and to contribution levels, and (4) the options allowed individuals to choose based on their preferences and goals (e.g., the preference to increase contributions to pursue the goal of a raise). As this example neatly shows, the construction of choice conditions and arrays allows one to nudge choices, and nudged choices increase the probability of targeted behaviors.

We now turn to the final construction of choice through gamification with the ubiquity of games and their role in capitalizing on and increasing

28. Thaler, Richard and Shlomo Benartzi. "Save More Tomorrow: Using Behavioral Economics to Increase Employee Saving." *Journal of Political Economy* 112, No. S1 (2004): S164-S187.

motivation. Researchers classify games as voluntary activities which are enjoyable and rewarding, provide clear feedback and evidence of successful action, promote goals, and highlight autonomous choice.[29] Gamification is the design of services and systems to incorporate features of games, most specifically their reward mechanisms that incline choice by increasing motivation and enjoyment through the inculcation of value. Consequently, to say that something is "gamified" means that it models game reward mechanisms to increase motivation.

Gamification is increasingly used within education, particularly in e-learning contexts where students are more disengaged and do not have the same level of in-person and social interaction or feedback from others as in traditional classroom contexts.[30] In many ways, gamification capitalizes on the very psychological processes that inculcate either virtue or vice. Games provide clear goals in terms of outcomes, which individuals can identify as either being in alignment with or opposed to their preferences. For example, knowing the goal of solitaire, I can decide if I want to engage in that game (which involves symbol matching, memory, and logic) or if I prefer to play a board game (that has more interaction with others). Games provide clear feedback, which functions as direction and correction. For example, if I rotate the shape in the wrong direction when playing a game of Tetris, it fails to fit the slot. This failure provides me not only with information about what I should do (direction), but also an immediate correction to guide my action for the next shape. And I have evidence of a successful action that demonstrates the efficacy of my agency, because each decision and outcome are clearly and directly linked within a short timeframe. Further, games are rewarding, both literally and psychologically. Through their structure for accruing points, badges, and levels, they provide literal rewards. They are also psychologically rewarding, in that they provide an escape from stresses and are highly enjoyable, as well as allowing us to amass perceived value and increase pleasure or satisfaction through the activity itself.

As we have seen throughout this section, we construct both situations and choice contexts to incline individuals toward desired choices and targeted behaviors, and to cultivate virtue through consistent action, practiced behaviors, and the development of intrinsic motivation. In the next section, we draw some insights into how these tools can be used to foster civility in online contexts.

---

29. Hamari, Juho, Jonna Koivisto, and Harri Sarsa. "Does Gamification Work? – A Literature Review of Empirical Studies on Gamification." *2014 47th Hawaii International Conference on System Sciences* (2014): 3025-3034; Koivisto, Jonna and Juho Hamari. "The Rise of Motivational Information Systems: A Review of Gamification Research." *International Journal of Information Management* 45 (2019): 191-210.

30. An, Yunjo. "Designing Effective Gamified Learning Experiences." *International Journal of Technology in Education* 3 (2020): 62-69.

INCULCATING CIVILITY IN ONLINE CONTEXTS

As we have seen, goal selection and automaticity, priming, nudges, in-group priming, and gamification are each powerful tools that incline ideas, interpretations and perceptions, choices, and behaviors. Clearly, a wholesale change to online platforms would be counterproductive, as users will not continue to use platforms that differ substantially from those with which they are currently familiar. But modifications to online platforms, which make use of some combination of the tools above, can maintain user experience yet yield incremental changes in how individuals interact with others. While not everyone will be civil in all instances and at all times, online platforms can encourage some choices and behaviors and discourage others. Over time and through continued opportunities to engage civilly with others, with constant reminders and models of civility, with continual corrective feedback, and with forms of positive reinforcements and motivational drivers, online platforms can foster civility as an intellectual virtue. Although technology is changing rapidly, we can sketch how each of the tools discussed above could be implemented within online platforms as a redesign of users' situational contexts and choice options.

One of the simplest tools to apply would be goal formation and priming. Upon the creation of an account, a user could be required to fill out or select the goals they have for using the platform. Given the existing services provided by social media platforms, those goals would likely include things such as conversation, interaction, forming connections, reestablishing lost friendships, sharing family information and memories, learning from information boards, networking in the community for causes (e.g., local resale or freecycling), etc. Although some individuals do create accounts to spread disinformation or sew discord as trolls, the vast majority of individuals sign up for social media accounts specifically to use the platforms to further their social networks. The simple activity of forming or identifying our goals serves to raise them to conscious accessibility, which also increases the probability of attendant and related concepts that help to guide our thoughts and behavior. If I select the goals of "forming connections" and "learning from information boards," I am more likely to entertain an unfamiliar idea charitability in order to maintain the connections I have developed. Bringing our goals to mind so that they help guide our behavior needs not only be something that occurs during account creation; rather, one's goals could be used as a prime during an annual renewal process. Each year, users could be asked to review and update their goal card (no differently than how users must routinely update current email and phone contact information, or review and agree to current privacy policies). Again, the repeated prime of reviewing one's goals serves to raise them to our conscious awareness for increased accessibility of use in our thoughts and behavior.

Other simple forms of primes that would be easy to apply would be daily reminders[31] as pop-up messages with checklists of parameters for acceptable and unacceptable behaviors. Rather than lofty mission statements of commitments to free speech, long and dry lists of rules buried in the depths of the platform, or community content policies which no one ever reads (unless informed that they have run afoul of them), a simple and clear checklist detailing acceptable and unacceptable behaviors for posts and responses serves as a constant reminder and prime to help guide action. Users could be required to either manually check off each item, or simply click a button to affirm that they have reviewed the checklist. Even if they fail to read the checklist each time, or merely click the button as a matter of habit, the content on the screen scene in their peripheral vision will still have a powerful effect on their non-conscious processing, functioning no differently than the powerful and additive effect of advertising on our preferences and choices.

Nudges are also easy to implement, both in simple and complex forms. Simple forms of nudges can provide reminders to interact with others and thereby maintain and increase social connections. Like the prompts generated by Gmail that note the number of days since a message was received and suggest a reply, users could be nudged to respond to others in social media private messages or in response to direct public posts. Clearly, implementing such a nudge would be an easy lift, as the technology already exists and is in active use in messaging applications. For platforms that use algorithms to track heated or inflammatory content via keywords or phrases,[32] this use of nudges would be particularly valuable to mitigate the voluntary dissolution of associations and to foster conditions for repeated exchange which are needed for accountability. While we might need a period to cool off after a heated interaction, a nudge to re-engage (along with a note of the number of days since the exchange) raises the probability for another exchange to occur. While there is no guarantee that a follow-up exchange will be less heated, the platform can foster repeated exchanges to maintain social connections. This simple design change presents users with altered choice arrays and encourages re-engagement rather than ghosting,

31. Sobolev refers to such smart feedback and reminders as "digital nudging*."* Sobolev, Michael. "Digital Nudging: Using Technology to Nudge for Good." *In Behavioral Science in the Wild*, eds. Nina Mazar and Dilip Soman (Toronto, Canada: University of Toronto Press, 2022), 292-299.

32. Computer engineers already track content such as hate speech, antagonistic language, offensive terms, and heated exchanges through algorithms that evaluate natural language processing and sentiment. Liu, Leyu, Xin Huang, Jianliang Xu, and Yunya Song. "Oasis: Online Analytic System for Incivility Detection and Sentiment Classification." *2019 International Conference on Data Mining Workshops (ICDMW)*, pp. 1098-1101. IEEE, 2019; D'Andrea, Alessia, Fernando Ferri, Patrizia Grifoni, and Tiziana Guzzo. "Approaches, Tools and Applications for Sentiment Analysis Implementation." *International Journal of Computer Applications* 125 (2015): 26-33; Watanabe, Hajime, Mondher Bouazizi, and Tomoaki Ohtsuki. "Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection." *IEEE Access* 6 (2018): 13825-13835.

muting, blocking contact, or leaving a social network.

An additional simple nudge to remind users of the parameters of acceptable and unacceptable behavior, as well as their social goals, would be to implement a "second step review" for all messages and posts. Simple forms of this nudge could require users to review their message content along with a pop-up query such as "Are you sure you want to send this message?" Unless messages achieve confirmation of both review and intent to send, they remain in draft status. This nudge creates an opportunity for reflection and review as part of the platform and communicative exchange. Simple forms of this nudge could be required prior to each message or post, but a more nuanced application could combine the nudge with algorithms that track heated content, targeted keywords, or patterns[33] and only prompt users for a "second step review" in cases of prior demonstrated incivility or where the probability for incivility has increased.[34]

More complex forms of priming could highlight social relationships through group membership and identity. Currently, social media platforms operationalize social connections as a quantified number rather than attending to the quality and kind of associations and social connections. Connections are measured in terms of approval or disapproval rankings (via likes, dislikes, thumbs up, angry emojis, etc.), or in terms of the sheer number of connections (number of friends, views, or responses to posts). In contrast, a more granular approach to social connection would highlight specific relations and commonalities between individuals. Such a design shift would quantify the degree of connection between individuals by tracking the number of shared groups, shared connections with other users, percentages of involvement/engagement within groups (e.g., number of posts in a knitting forum), etc. By tracking social connection through the more nuanced presentation of group and identity similarity, social media platforms can build on our psychological proclivity to treat those who are

33. The app BullyBlocker integrates with social media platforms such as Facebook to detect and track bullying patterns and frequency, creating a measure for bullying. Once bullying reaches a particular level, the app notifies parents so that they can intervene and provide additional support. Silva, Yasin, Christopher Rich, and Deborah Hall. "BullyBlocker: Towards the Identification of Cyberbullying in Social Networking Sites." *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (2016): 1377-1379. For more information on the app's development, grant funding and additional research applications, and integration with various social media companies, see https://news.asu.edu/20170920-solutions-asu-team-takes-cyberbullying-app-public .

34. Such technology is already beginning to be used for such a purpose. The app ReThink uses a specialized algorithm to detect bullying content and then prompts users to pause and reconsider sending a message before actually doing so. The app was developed by student Trisha Prabhu. For more information on the ReThink app, visit: https://www.rethinkwords.com/whatisrethink . Similarly, Instagram identifies content that is similar to previously flagged content along with a message notifying the user and providing the option to proceed with or delete the post. Sullivan and Reiner argue that allowing such user reflection (rather than automatic deletion from the platform) preserves agency. Sullivan Laura and Peter Reiner. "Ethics in the Digital Era: Nothing New?" *IT Professional* 22 (2020): 39-42.

members of our in-group favorably.[35] Displaying these new metrics on user pages provides a simple and immediate form of in-group priming. Making this information available and salient to users constructs a situation in which individuals are more likely to consider and evaluate ideas by assuming that others are trustworthy, justified, and reasonable. Slightly more complex and direct in-group priming could be accomplished through viewable "contact cards" that appear during interactions. When sending a private message or responding to a public post, a dialog box could expand and include brief details about the other party using information from the new metrics, such as shared group memberships or interests. More advanced and targeted in-group priming could capitalize on using content algorithms to search for specific shared interests and connections between two parties. Although constant information about shared characteristics with someone else might be a distraction from the regular user experience, algorithms tracking the prior occurrence of or the probabilistic increase of uncivil exchange could again be used to trigger a pop-up reminder with information about the degree of connection one has with another. Often, a simple reminder of the connection and the importance of that person to our social circles and our lives alters both how we interpret content and how we respond.

Using AI text generators, more complex forms of nudges can provide both direction and correction through models of civil communication.[36] Open AI's ChatGPT[37] has demonstrated an extraordinary ability to craft well-written, creative, and informed content that duplicates the linguistic and communicative quality of natural language users. The quality and speed with which AI text generators can create content make them an invaluable tool for the creation of several "content contrast" nudges.[38] The first is

35. Xiang, Rongjing, Jennifer Neville, and Monica Rogati. "Modeling Relationship Strength in Online Social Networks." In *Proceedings of the 19th International Conference on World Wide Web*, pp. 981-990. 2010. See Golbeck's article for an example of the way that more nuanced profile analysis of similarity affects user's judgments of trust. Golbeck, Jennifer. "Trust and Nuanced Profile Similarity in Online Social Networks." *ACM Transactions on the Web (TWEB)* 3 (2009): 1–33.

36. The Angry Uncle chatbot is an early version of a contrastive technology to provide models of civil communication. Sincere thanks to Lisa Schirch for sharing this resource. See Tamerius, Karin. "How to Have a Conversation With Your Angry Uncle Over the Holidays." *The New York Times.* https://www.nytimes.com/interactive/2018/11/18/opinion/thanksgiving-family-argue-chat-bot.html.

37. For more information or to create an account to use the chatbot, see: "Introducing ChatGPT." OpenAI, https://openai.com/blog/chatgpt.

38. Researchers have developed a variety of "counter-narrative" strategies in which natural language generation (NLG) models (such as OpenAI's ChatGPT) mimic a communicative response by a human to "counter" (challenge, address, or respond) posts containing harmful, inflammatory, or hate speech content. Carla Schieb and Mike Preuss. "Governing Hate Speech by Means of Counterspeech on Facebook." 66th ICA Annual Conference (2016): 1–23; Mathew, Binny, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. "Thou Shalt Not Hate: Countering Online Hate Speech." *Proceedings of the International AAAI Conference on Web and Social Media* 13 (2019): 369–380; Ashida, Mana and Mamoru Komachi. "Towards Automatic Generation of Messages Countering Online Hate Speech and Microaggressions." *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), Association for Computational Linguistics* (2022): 11–23. While this research is promising, it is functionally equivalent to interacting with a real "agent" that counters one's claims in an argument. (Indeed, NLG models are

something we might call the "Charity Generator." Suppose that a user receives a message or post with either moral or political content from another. When the user clicks on the content to reply, the Charity Generator instantly crafts an alternate version of the message that removes inflammatory language, derogatory terms, or trigger words, adds component steps of an argument, premises, or reasons, and identifies a clear conclusion that follows from the presented claims and reasons. In short, the AI text generator crafts the best possible interpretation of the content that a reasonable, informed, and civil writer would draft. The original and AI-generated versions are presented side by side for review, and a pop-up message queries the user: "Does this revised content capture B's intended message?" If the user clicks "Yes," then the platform minimizes the original message box and leaves the altered message for review as the user drafts a response. If the user clicks "No," the platform minimizes the alternative message box and leaves the original message in place. The choice between the two versions serves as a nudge for the user to think more charitably and respond more civilly. Most importantly, this nudged choice maximizes the possibility for civil exchange, as the user drafts a reply based on the most charitable version of the message. Even if the user rejects the altered message as reflecting B's intent, the AI-generated version still functions as a model—and hence, a prime—of civil communication in contrast to the actual message.[39]

The second form of "content contrast" nudge to consider is something we might call the "AI Garbler." Triggered by either moral and political content or algorithms tracking the probability of uncivil exchange, this AI text generator would instantly craft an alternate version of one's own message. The platform would display both versions to the user prior to sending or posting with the query: "Does this revised wording better capture your intended message?" But whereas the previous AI text generator develops more charitable versions, the Garbler deliberately makes messages less clear with irrelevant examples, vague phrases, ill-formed sentences, and the use of non-words. Obviously, the user will not select the garbled version. Upon selecting "No," the user can return to and revise the original message. If the revised message avoids inflammatory language, the message can be sent or posted. But if it trips the algorithmic alarms once again, the Garbler will instantly craft an alternative version of the new message and the process

designed to be functionally equivalent in generating the natural language responses of humans.) While such counter-narrative methods using NLG models may succeed in changing someone's mind for a specific harmful claim (e.g., racist), the success rate is likely to be no higher than that achieved by human agents trained as content moderators. The main benefit touted for these strategies is a greater ability to achieve scale given the limited number of trained human content moderators on social media platforms.

39. While posts by NLG models do provide reasons and evidence in response to a user's posts, and thereby model good reasoning, they do not induce a user to interpret another's posts with greater charity. In contrast, one of the primary benefits of the Charity Generator is that it provides users with practice in interpretation through considering and evaluating multiple meanings behind literal text.

repeats. This contrast of messages and the user's revision of his/her own content serves both as a nudge and as a form of user-directed training to draft better and less inflammatory content in the first place. To avoid negatively impacting the user experience, platforms could allow users to post their content without revision, but link audience size and reach to the number of revisions. Users who decline to revise their content earn a limited audience, while users who revise content in response to the Garbler have no restrictions on their audience. In both cases, users make choices based on their preferences and the platform design promotes civility.[40]

Gamification to promote civility can be incorporated in many ways within social media platforms. To gamify an activity, designers need to provide clear links between choices and outcomes as well as clear rewards for selected achievements. Suppose platform designers decide to pursue goal formation and priming as a tool to promote civility. Designers could offer points for the number of times per month that users review their goals, create benchmarks and levels toward goal achievement, and have algorithms award points for civil exchanges based on the difficulty level and percentage of moral and political content. Users could monitor their overall achievement via a simple sidebar scorecard. Providing accessible information and a way to measure and track achievement motivates users by making voluntary choices clear, actionable, and rewarding.

Each of the above tools could be gamified in isolation, but a combination of them could motivate the development of civility in powerful ways. Rather than a simple sidebar scorecard to track a single factor, designers could create a user badge or "civility label." Product labels contain information about a variety of nutritional factors, such as carbohydrates, fats, calories, and ingredients. Nisolo's new sustainability label[41] provides information about the various factors important for increasing sustainability along with achievement scores and ratings for each factor. Consumers can use this clear and accessible information to guide their purchases through informed choice. Similarly, a user "civility label" would present information about the factors important for civil discourse and reflect an individual's achievement of civility. The civility label could be part of each user's login screen or a constant menu of items that provides clear and accessible information to each user about his/her personal performance.

40. Whereas the NLG counter-narrative approaches mentioned in a previous footnote provide a reactive response to a user's post, the Garbler Generator teaches the user to modify the content of their messages in the drafting process. Users learn to adjust their content based on the immediate feedback of text garbling, and are able to revise their content prior to a post. This pre-posting revision process allows users to reconsider the content of their messages in a less emotionally charged context than receiving a counter. One of the primary benefits of the Garbler Generator is that it provides users with practice in reflection on and revision of content.

41. For more information, see: "Nisolo launches the Sustainability Facts Label." Make Fashion Better. January 17, 2022. https://www.makefashionbetter.com/blog/nisolo-launches-the-sustainability-facts-label.

Social media designers could select factors they desire to promote such as goal creation, low percentages of personal attacks, low percentages of inflammatory language, high percentages of selections to revise content (via the Charity Generator), etc. Providing users with accessible information, metrics to mark achievement, and clear choice and outcome pairings so users can make improvable action capitalizes on the design of gamified systems. Users are competing with themselves to increase their scores, maximize their achievement levels, and earn the highest rankings demonstrating civility.

As should be clear from the brief sketches above, there are many tools that could be used in online contexts to increase the probability of civil exchanges and foster civility. A side benefit of implementing these tools is that the additional reminders, primes, or nudges often require some additional step or action of review. This additional step serves to slow communication and limit reactive emotional responses (which cyclically helps the nonconscious and conscious situational constructions to be more effective). Goal selection and automaticity, reminders, primes, nudges, and gamification shape discrete choices and individual actions in subtle yet powerful ways.

CONCLUSION

While some versions of these tools may have been deployed as ways to reduce vitriolic messages, responses, and postings by social media platforms such as Reddit, Facebook, TikTok, and Twitter, the approach has not been one of fostering civility but of creating policy for content governance. The emphasis on the legal and institutional/organizational development of policy has distracted from the consideration of whether and how those policies affect individual conditions of choice. Worse, the focus on policy rather than the structural conditions of choice and agency has allowed incivility to spread. Discrete moments of engagement provide opportunities to train and shape actions and choices, whether negatively or positively for vices or virtues. When situations are not designed to foster intellectual virtue, they can easily promote intellectual vices by co-opting the same developmental processes. This shift to the psychological level and a focus on the intellectual virtue of civility yields powerful insights for technology policies and designs aimed at promoting better online discourse and strengthening democratic capacities. It not only offers a deeper understanding of the spread of incivility, but also offers direction for how we can arrest the problem.

This shift and reconceptualization of the problem allow us to see the new task. We need to develop not just new tech policies, but also new designs for online platforms that actively foster the development of civility as an intellectual virtue. Prohibitive policies do not provide informative and

nuanced guidance for action, and what users need is the guidance of civility. To cultivate intellectual virtues such as civility, we must attend to both the content of what individuals learn as well as the thousands of discrete experiences which form their collective training and learning grounds. Social media is an immersive experience, and each interaction and exchange is an educational opportunity within a process of lifelong learning and development. We must only decide what we—as policymakers, designers, and citizens—choose to foster as a possibility and as a responsibility.