**Comments on Saray Ayala, "Explaining Injustice in Speech: Individualistic vs. Structural Explanation," Minds Online 2015**

Alex Madva

Critics of individualistic approaches to discrimination argue that this systemic social problem stems not from the irrationality, selfishness, or prejudice of individual actors, but from underlying structural-institutional forces. Philosophers often voice this criticism of individualism, but less often flesh it out in detail, or offer a substantive, alternative conception of discrimination in anti-individualist terms. Saray Ayala's rich paper goes beyond hand-wavy gestures to "think more about the social and less about the individual," and makes the substantive claim that persisting patterns of injustice in discourse depend primarily on shared social *conventions and norms* rather than biased mental states. Ayala also points toward directions for future research, calling on philosophers and scientists to explore in greater depth the socially situated and embodied aspects of speech injustice.

Speech injustice occurs, roughly, when a speech act receives inappropriate uptake from others due to the speaker's social identity, in a way that reinforces broader patterns of social disadvantage. Ayala's example is a non-native speaker who contributes to the Q&A after a talk, "but nobody engages with it." Then a native speaker effectively repeats the point, and animated debate ensues. Speech injustice is a descendant of Rae Langton's (1993) illocutionary silencing, Miranda Fricker's (2007) testimonial injustice, and Rebecca Kukla's (2014) discursive injustice.

Ayala contrasts two potential explanations of speech injustice. One, predominant in recent philosophy and social psychology, appeals to biased minds. Fricker and Kukla trace the injustice in these cases to hearers' prejudices, which in turn reflect pervasive social attitudes. Kukla (2014, 447) locates the proximate cause of discursive injustice in "subterranean assumptions and habits concerning gender, embodiment, and power that are hard to articulate."[1] I'll call this emphasis on bias a B-explanation. Ayala proposes instead that speech injustice is more a matter of conventions and norms. To use Fricker's example from *The Talented Mr. Ripley*, when Herbert Greenleaf says, "Marge, there's female intuition, and then there are facts," perhaps he is not so much prejudiced against women as he is employing a shared (unjust) convention of discounting women's testimony or rationality. In Kukla's terms, there might be a convention of treating women's testimony as an entreaty to join the conversation, rather than a direct contribution to it. I'll call these C-explanations.

Before weighing their relative merits, it's worth emphasizing a central feature these explanations share. Both B- and C-explanations appeal to unconscious, or at least tacit and not-fully-articulated, phenomena. People might harbor a certain bias, or follow a certain convention, and not be fully aware of it. They might sincerely disavow the bias/convention, and mistakenly deny that their behavior ever reflects it. While implicit biases are often glossed as completely unconscious, evidence suggests that people are aware of them *qua* "gut feelings" (e.g., Ranganath et al. 2008). Frequently, however, we don't explicitly *notice* our biased gut feelings, or appreciate how a given bias is influencing us in a specific context. The same is likely true of tacit conventions. Even if we are not explicitly aware that our behavior follows some shared, rule-governed pattern, we might feel an inchoate sense of discomfort when our behavior falls out

[1] I'm not sure this characterization is fair to Kukla (or Fricker), who does not deny the existence of "politically unfortunate" discriminatory discursive conventions (448). She's just not focused on *unjust* conventions in that paper. She is analyzing cases where speech acts meet all relevant criteria according to established conventions, but the performative force is distorted due to biases against the speaker's social identity. Ayala also suggests that Kukla claims that people must be aware of a pattern in order for it to count as a convention, but I don't see this claim. Kukla (2014) only says it needs to be stable and "sufficiently regular." Finally, while Kukla (like Fricker) traces the *injustice* in question to hearers' biases, it is not true that she "leaves aside all… structural constraints." Kukla offers a rich account of the social-structural dynamics in these cases, which Ayala draws on. Kukla's paper exemplifies how an appreciation of the role of bias can be integrated into a complex account of social dynamics. I also wonder whether it might be more apt to say Ayala disagrees with Kukla about the nature or paradigm cases of discursive injustice, rather than to coin the new term "speech injustice."

of synch with that pattern. A standard example might be norms guiding how far away we stand from interlocutors, which are usually noticed only when we come across, e.g., a close-talker who violates them. Both B- and C-explanations posit these peripherally-conscious phenomena, and thereby allow for conflicts between sincere avowals of egalitarianism and unwittingly discriminatory behavior.

It's also worth disambiguating a few different claims that might be at play. One relatively uncontroversial claim:

**Moderate Anti-Individualism**: it is not the case that bias explains *all* cases (or components) of speech injustice.

Ayala writes, for example: "A complete account of hearers' situatedness in the social reality must consider not only their attitudes and beliefs, even if they are a result of their exposure to social factors, but also the social factors themselves that, at that moment and in that space, are guiding their interpretation of speaker's words." Attitudes are one part of a broader, complex story. I think few defenders of B-explanations would deny this, although the sheer amount of time and words many of us (myself included) have spent focusing exclusively on implicit bias (e.g., as it pertains to individual moral responsibility) rather than on social structures may suggest that we have devoted excessive attention to the individualistic pieces of the puzzle. (We may ourselves be subject to kind of attributional bias or convention, dwelling on those causes of injustice located in individual minds and overlooking situations and structures.)

Another claim, which makes an important contribution to our understanding of speech injustice and other microinequities:

**Moderate Conventionalism**: conventions explain a significant subset of cases (or components) of speech injustice.

I am persuaded by Ayala that the role of conventions in speech injustice has been underexplored. At times, however, Ayala gestures toward two more controversial claims:

**Strong Anti-Individualism**: bias explains *no* (philosophically or politically interesting?) cases (or components) of speech injustice.

For example, Ayala writes that B-explanations of injustice "could appropriately account for the phenomenon only in a society where interactions among individuals are *not* governed by unjust conventions." Such claims suggest that as long as unjust conventions are in place, appeals to bias make no further contribution to explaining, predicting, or understanding speech injustice (see esp. §2.2 on norm-conforming behavior).

Another controversial claim:

**Strong Conventionalism**: conventions explain *all* cases (or components) of speech injustice.

For example, Ayala seems to say without qualification that speech injustice "is governed by conventions."

Either of these stronger claims is, I think, a pretty hard sell.  Couldn't there be a) some cases where relatively unbiased minds follow biased conventions, b) some cases where biased minds flout unbiased conventions, and then c) "perfect storm" cases where biased minds follow biased conventions?  The conjunction of Moderate Anti-Individualism and Moderate Conventionalism seems on firmer ground, and I agree with Ayala that, "Effective interventions against speech injustice will have to take into account both individuals' minds and the conventions that constrain individuals' behavior" (6).

Now let's wade into the substantive virtues and vices of B- vs. C-explanations.

**Stability.** One central advantage that Ayala cites for C-explanations is stability. Conventions are more stable across times, places, and people than are mere mental states, which can vary dramatically across contexts.  B-explanations are unstable in the counterfactual sense that subtle differences in people's minds and contexts would make a given B-explanation inappropriate.  Structural explanations "abstract away from certain features of the individual and allow for 'inessential perturbations' in the sequence of events" (Haslanger 2015, citing Garfinkel 1981).

I wonder, however, whether the relative instability of B-explanations is a virtue rather than a vice.  Speech injustices are stable in the sense that they continue despite dramatic declines in overtly discriminatory beliefs, norms, laws, etc.  But are they stable in the further sense that they occur regardless of subtle perturbations in minds and contexts?  It is not that *every* comment made by non-native speakers at colloquia is ignored.  If the phenomena were so regular and predictable, their existence and relevance might be less controversial.

One reason for continuing skepticism about implicit bias, microinequities, cumulative (dis)advantage, and the general persistence of subtle and informal discrimination is that effect sizes in empirical investigations are often small (Oswald et al. 2013). Studies on NBA referees (Price and Wolfers 2010) and MLB umpires (Parsons et al. 2011) found patterns of racial bias in calling fouls and strikes, especially in "borderline" cases, but vast quantities of data needed to be analyzed to detect bias. The patterns of discrimination were real, and arguably made a difference to game outcomes, but eluded detection via ordinary spectating. These patterns are stable in that they endure over time, affect many individuals, and crop up in many contexts, but are radically unstable in that one could not predict precisely when they'll occur or say with confidence that any particular decision was affected by race. The vast majority of decisions were accurate, making it difficult to think that any referees were, consciously or unconsciously, following a racist foul-calling convention. It seems more natural to say that they were slightly biased.

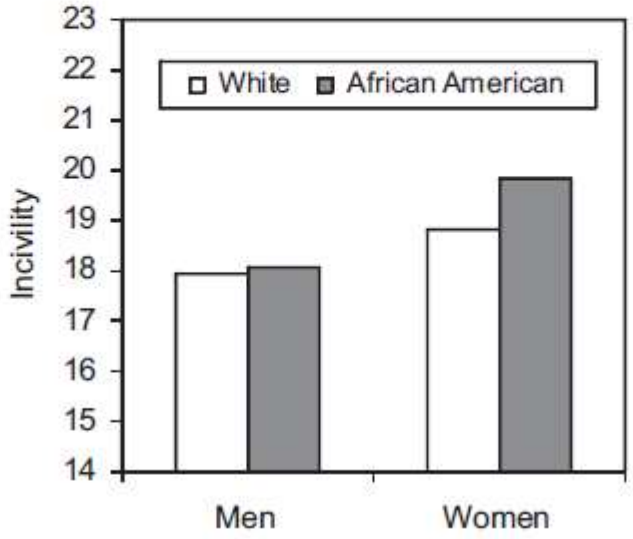
More relevant to speech injustice, Lilia Cortina and colleagues (2011) found that women, and especially women of color, tended to report experiencing more incivility in the workplace than do men.[2] In many cases, the incivility consisted not in overt harassment but in generic forms of rudeness, such as speaking condescendingly or interrupting a colleague. Cortina also found that women's disproportionate experiences of incivility made them more likely to intend to quit, reinforcing existing evidence that experiencing a work environment as hostile leads one to quit (e.g., Sims et al. 2005).[3] Again, effect sizes were not overwhelmingly large. Participants reported how frequently, on a scale of 1 (never) to 5 (very often), they experienced each of 10 types of incivility from supervisors or co-workers, including "Paid little attention to your statements or showed little interest in your opinions" and "Doubted your judgment on a matter over which you had responsibility." A total score of 10 would mean never experiencing any of

[2] The following summary is adapted from a manuscript of mine on implicit bias and moral responsibility.

[3] One sort of anti-individualist response to studies like Cortina's would be to suggest that structural factors (e.g., the glass ceiling, race/gender wage gaps, lack of affordable daycare, gender disparities in parental leave, housing segregation, poor public transportation, etc.) are more predictive of race and gender disparities in employee turnover than speech injustice, but Ayala seeks to explain speech injustice *per se* in structural terms.

these incivilities and 50 would mean very often experiencing all 10 types. In one study on the US military, average scores of reported experiences of incivility were:

**Figure 1**
**Estimated Marginal Means for Gender-by-Race Effect on Incivility**



White men: 17.95

African American men: 18.07

White women: 18.83

African American women: 19.85

These differences are both real and really subtle. It's hard to look at them and think folks are generally following a full-blooded convention, conscious or unconscious, of uncivil treatment toward black women.

Such microinequities may, then, be highly unstable in that their occurrence is contingent on myriad normatively irrelevant factors and the passing whims of individual minds. For example, Forgas (2011) manipulated participants' moods before they read a philosophy essay either written by “a middle-aged bearded man in a suit with spectacles” or “a young woman with frizzy hair wearing a t-shirt.”  Forgas found that participants in a good mood relied on stereotypes and “gut feelings,” and evaluated the older man’s essay, competence, and likeability much more highly than the young woman’s.  However, being in a bad mood induced more vigilant, attentive thinking, and reduced this age/gender bias to statistical insignificance.

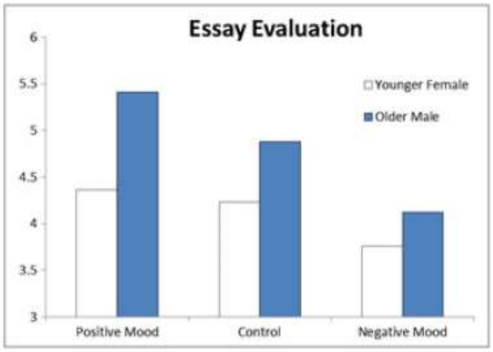


Figure 1.  Mood moderates halo effects on the evaluation of an essay: positive mood increased and negative mood eliminated the halo effect

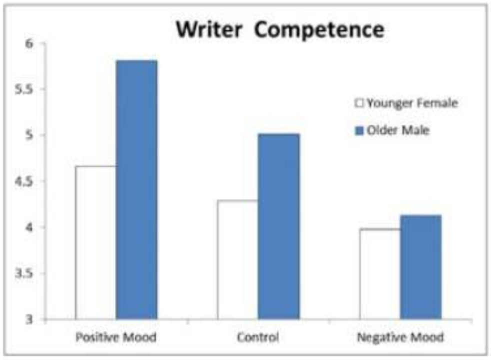


Figure 2.   The interaction of mood and halo effects on judgments of the competence of a writer: positive mood increases and negative mood eliminates halo effects

Again, fickle features of minds and contexts made the difference between egalitarian vs. discriminatory treatment.  By abstracting away from these perturbations, C-explanations may significantly *overpredict* the quantity and regularity of discrimination.  One way to save C-explanations might be to make them more fine-grained by indexing them to mental states and contexts, e.g., a norm of going with your gut when in a good mood.

In any case, the cross-contextual variability of contemporary discrimination may be closely related to one of its profound harms.  The incivility reported by Cortina’s participants is deeply ambiguous: just about *everybody* interrupts and gets interrupted *sometimes*, so it is difficult to identify any particular instance of interruption as expressive of bias, as opposed to, say, misplaced enthusiasm.  But this ambiguity itself is harmful.  Members of stigmatized groups sometimes find ambiguous but potentially biased behavior more unsettling and taxing than outright discrimination (Salvatore and Shelton 2007; Sue et al. 2007).  Discrimination might sometimes be easier—to cope with and to combat—if people "knew their enemy" was stable social conventions (or explicit prejudice) rather than inconsistently and ambiguously biased minds.

**Cases more amenable to C-explanations?** I believe some cases are better explained in terms of conventions, or at least by psychological or structural phenomena we have not yet found ways to measure. Take Rooth and colleagues' findings about discriminatory hiring practices in Sweden toward Arab-Muslims and obese individuals, e.g., identical résumés with Swedish names were three times more likely to get callbacks for interviews than résumés with Arab-Muslim names. Implicit measures predicted hiring discrimination over and above explicit measures, but the truth is that neither implicit nor explicit measures predicted much of the real-world discrimination. In one study, a full 58% of employers openly reported a preference for hiring thin people, but this explicit preference predicted *none* of the real-world discrimination. Maybe conventions/norms can better explain or predict patterns of discrimination that are predicted less well by implicit or explicit bias.

If so, this raises further questions. For conventions to be more than placeholders standing for "whatever it is, besides bias, that explains these patterns," it would be great to have clearer ways of measuring or detecting them. Under what conditions are we entitled to posit a convention or norm? One criterion might be that people report following the convention, but this won't apply for tacit discriminatory conventions. Another criterion regards how regular the behavior is. Another might be if people experience a sense of norm violation when the discrimination does *not* occur. That is, if a hearer *did* engage with the non-native speaker's insightful comment, might others feel irritated that this (by their lights) unperceptive or incomprehensible comment from an outsider was being treated as worthy of discussion? Might members of the audience more vigorously wave their hands to get their own “more pressing” question addressed, or otherwise try to intervene and steer the conversation away from what they take to be a sidetracking non-issue? Perhaps we could build here on Holton (2010) and Uttich and Lombrozo’s (2010) interpretations of Knobe’s side-effect effect. They argue that we are more likely to judge actions to be intentional if they violate a norm than if they conform to it. If people are more likely to think that the hearer who engages the non-native speaker does so intentionally than to think that the hearer who ignores the non-native speaker does so intentionally, then the latter speech injustice might be a full-blooded norm. The parallel prediction (a harder and more tantalizing test case, I suspect) would be a greater likelihood of attributing intention to hearers who ignore rather than engage high-status insiders’ comments. One potential snag: it seems essential to Holton’s account that the norm violation is *knowing*, that all parties are fully aware of the norm and the violator knowingly disregards it, but such explicit awareness is not involved in Ayala’s paradigm cases.

**Complementarity.** As Ayala sometimes emphasizes, these explanations can complement each other rather than compete. One proposal for joining the two is roughly this: the cognitive-affective-motor dispositions that we refer to as “implicit biases” *just are* the psychological constructs that track these oft-unspoken, oft-unjust discursive norms. Michael Brownstein and I (2012a,b) argued for a claim along these lines. Whereas implicit biases are often depicted as arational associations, we argued that they are highly flexible and sensitive to certain norms. So I sympathize with Ayala’s claim that many cases of speech injustice involve “perfectly skilled listeners who are appropriately applying the conventions operative in their communities.” The skill and norm-sensitivity of implicit biases is often overlooked by philosophers and psychologists, and more research on the socially situated and embodied aspects of these phenomena is clearly necessary.

**Framing the debate.** Although we were trying to capture an idea similar to Ayala's, our response was to propose a revision in conceptualizing implicit biases (a revision in psychological theory), rather than a redirection of attention away from the psychological and toward the social. This leads to my final question. I worry that, in some contexts, framing the issue in terms of individual *versus* structural explanations can be misleading. A complete explanation should, I believe, make reference to individual as well as social, structural, and institutional phenomena, although which of these aspects is more salient to our local explanatory aims will of course vary. Boo reductive, methodological individualism! Boo reductive social explanations that gloss over psychology altogether! Ayala at times seems to agree.

Often, the upshot of criticisms of individualism may not be that we should switch from one type of explanation to another (psychological to structural), but that we shift our attention *within* each of these domains. For example, Ayala's argument calls *inter alia* for a shift within psychology: from the study of irrational unconscious biases to the study of skillful norm-following. Understanding speech injustice requires, in part, understanding how individuals learn, follow, sustain, and revise conventions and norms, perhaps in ways that elude introspective awareness. The relevant psychological dispositions are not just stereotypical associations passively imprinted on the mind by "mass media," but reflect a skillful ability to acquire and seamlessly follow norms. There is a parallel shift in social science: from conceiving discrimination solely in terms of the aggregate of individuals' preferences and beliefs, to the shared conventions and social structures that constrain behavioral options.

**References**

Agerström, J., & Rooth, D. O. (2011). The role of automatic obesity stereotypes in real hiring discrimination. *Journal of Applied Psychology*, *96*(4), 790.

Brownstein, M.S., and Madva, A.M. 2012a: Ethical Automaticity. *Philosophy of the Social Sciences*, doi: 10.1177/0048393111426402.

Brownstein, M.S., and Madva, A.M. 2012b: The Normativity of Automaticity. *Mind & Language* 27 (4), 410-434.

Cortina, L.M., Kabat Farr, D., Leskinen, E., Huerta, M. and Magley, V.J. 2011: Selective incivility as modern discrimination in organizations: Evidence and impact. *Journal of Management*.

Forgas, J. (2011). She just doesn't look like a philosopher. . .? Affective influences on the halo effect in impression formation. *European Journal of Social Psychology*, 41, 812–817 (2011)

Fricker, M. 2007. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford: Oxford University Press.

Garfinkel, A. (1981). *Forms of explanation: Rethinking the questions in social theory*. New Haven: Yale University Press.

Haslanger, S. 2015a. What is a (Social) Structural Explanation? *Philosophical Studies* 172.

Holton, R. 2010. Norms and the Knobe Effect. *Analysis* 70 (3):417-424.

Kukla, R. 2014. Performative Force, Convention, and Discursive Injustice. *Hypatia* 29 (2)

Langton, R.1993. Speech Acts and Unspeakable Acts, *Philosophy & Public Affairs* 22: 293-330.

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting Ethnic and Racial Discrimination: A Meta-Analysis of IAT Criterion Studies. *Journal of Personality and Social Psychology*. doi: 10.1037/a0032734

Parsons, C.A., Sulaeman, J., Yates, M.C., & Hamermesh, D.S. (2011). Strike Three: Discrimination, Incentives, and Evaluation. *American Economic Review*, 101: 1410–1435.

Price, J., & Wolfers, J. (2010). Racial Discrimination among NBA Referees. *The Quarterly Journal of Economics*, 125 (4) 1859-1887.

Ranganath, K. A., Smith, C. T., and Nosek, B. A. (2008). Distinguishing automatic and controlled components of attitudes from direct and indirect measurement methods. *Journal of Experimental Social Psychology*,*44*(2), 386-396.

Rooth, D. O. (2010). Automatic associations and discrimination in hiring: Real world evidence. *Labour Economics*, *17*(3), 523-534.

Salvatore, J., and Shelton, J.N. 2007: Cognitive costs to exposure to racial prejudice. *Psychological Science*, 18, 810-815.

Sims, C., Drasgow, F., and Fitzgerald, L. 2005: The effects of sexual harassment on turnover in the military: time-dependent modeling. *Journal of Applied Psychology* 90, 1141-1152.

Sue, D.W., Capodilupo, C.M., Torino, G.C., Bucceri, J.M., Holder, A.M.B., Nadal, K.L., and Esquilin, M. 2007: Racial microaggressions in everyday life: Implications for clinical practice. *American Psychologist* 62, 271-286.

Uttich, K. & Lombrozo, T. 2010. Norms inform mental state ascriptions: a rational explanation for the side-effect effect. *Cognition* 116:87-100.