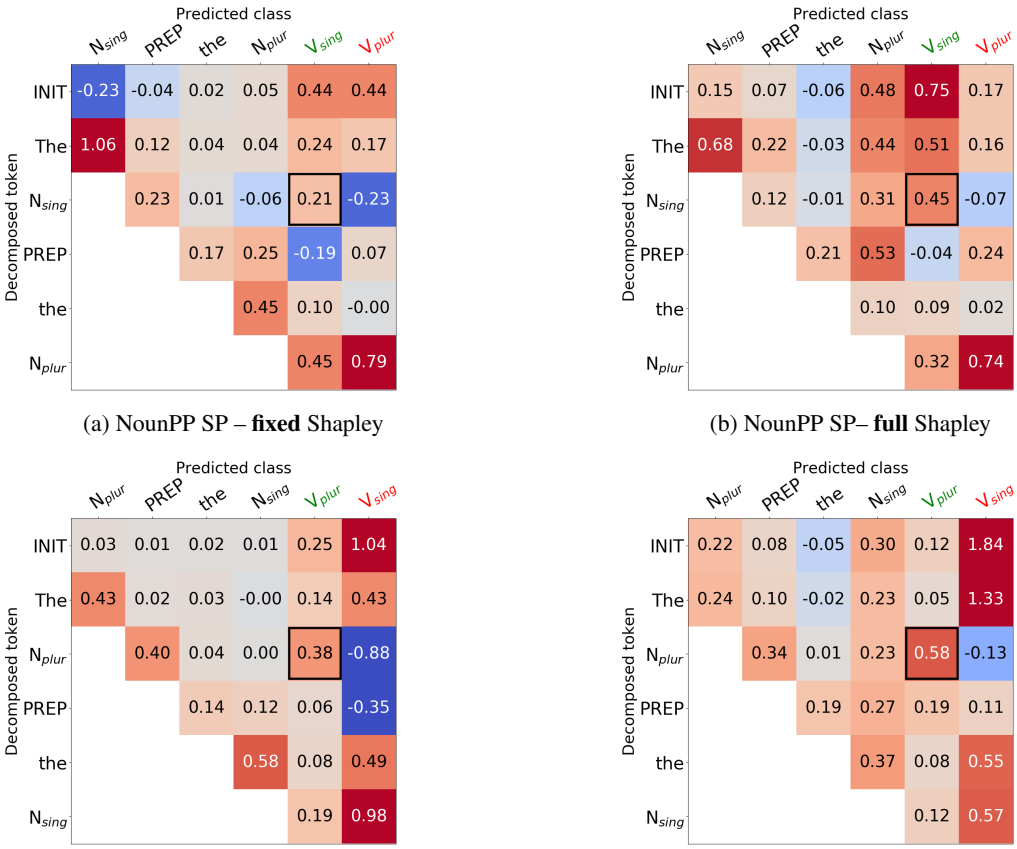
# A Fixed Shapley results – number agreement




(a) NounPP SP – **fixed** Shapley

(b) NounPP SP– **full** Shapley

(c) NounPP PS – **fixed** Shapley

(d) NounPP PS – **full** Shapley

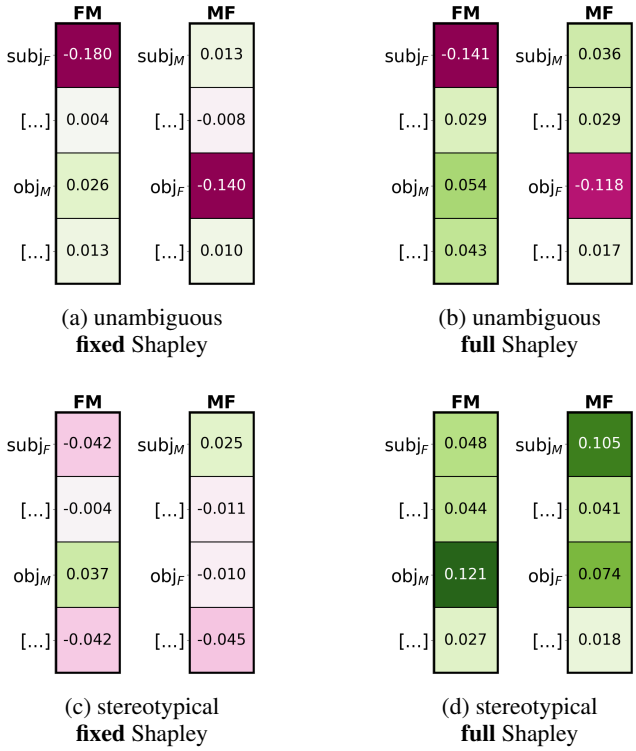
Figure 4: Results of Figure 2, for both Shapley computations. Note how the fixed Shapley results generally lead to lower term contributions, as these are more prominently assigned to the intercept terms instead.

| | | GCD – **fixed** Shapley | | | |
|---|---|---|---|---|---|
| **NA Task** | **C** | FULL | IN | INTERCEPT* | ¬INTERCEPT |
| Simple | S | 100 | 100 | 100 | 7.7 |
| Simple | P | 100 | 100 | 7.3 | 65.7 |
| nounPP | SS | 99.2 | 91.2 | 100 | 14.8 |
| nounPP | SP | 87.2 | 91.7 | 100 | 14.3 |
| nounPP | PS | 92.0 | 100 | 0 | 82.7 |
| nounPP | PP | 99.0 | 99.8 | 0.5 | 81.0 |
| namePP | SS | 99.3 | 91.2 | 100 | 12.4 |
| namePP | PS | 68.9 | 99.8 | 0 | 82.0 |

| | | GCD – **full** Shapley | | | |
|---|---|---|---|---|---|
| **Task** | **C** | FULL | IN | INTERCEPT* | ¬INTERCEPT |
| Simple | S | 100 | 73.3 (91.3) | 97.3 (100) | 69.7 (86.3) |
| Simple | P | 100 | 100 (100) | 32.7 (7.7) | 100 (100) |
| nounPP | SS | 99.2 | 93.0 (99.7) | 99.8 (99.8) | 72.7 (88.7) |
| nounPP | SP | 87.2 | 90.3 (99.3) | 98.8 (99.8) | 60.5 (83.5) |
| nounPP | PS | 92.0 | 100 (100) | 0.0 (0.0) | 100 (100) |
| nounPP | PP | 99.0 | 100 (99.3) | 7.0 (0.5) | 99.8 (100) |
| namePP | SS | 99.3 | 97.7 (91.3) | 99.4 (100) | 76.2 (90.9) |
| namePP | PS | 68.9 | 98.3 (98.2) | 1.3 (0.0) | 99.9 (99.9) |

Table 3: Results of Table 1, for both Shapley computations. The main difference here lies in the ¬INTERCEPT case: for the fixed Shapley this case leads to a much starker decrease. The pattern, however, remains unaltered: the singular conditions depend much stronger on the intercepts than the plural conditions for both the Shapley computations.

## B Fixed Shapley results – pronoun resolution




(a) unambiguous **fixed** Shapley

(b) unambiguous **full** Shapley

(c) stereotypical **fixed** Shapley

(d) stereotypical **full** Shapley

Figure 5: Results of Figure 3, for both Shapley computations. The pattern remains the same, although the full Shapley case highlights a stronger default male bias that is encoded in the non-gendered sub-phrases.

| | GCD – **fixed** Shapley | | | |
|---|---|---|---|---|
| **C** | FULL | SUBJECT | OBJECT | INTERCEPT |
| MM | 100 | 100 | 100 | 100 |
| MF | 58.6 | 100 | 31.2 | 100 |
| FM | 37.0 | 6.2 | 100 | 100 |
| FF | 1.2 | 50.0 | 73.6 | 100 |

(a) *%he>she*, unambiguous referents

| | GCD – **full** Shapley | | | |
|---|---|---|---|---|
| **C** | FULL | SUBJECT | OBJECT | INTERCEPT* |
| MM | 100 | 100 (93.2) | 100 (97.8) | 100 (93.2) |
| MF | 58.6 | 100 (86.4) | 47.2 (0.8) | 100 (96.0) |
| FM | 37.0 | 29.2 (0.6) | 100 (97.2) | 100 (98.0) |
| FF | 1.2 | 77.2 (0.8) | 88.8 (1.2) | 100 (92.2) |

(b) *%he>she*, unambiguous referents

| | GCD – **fixed** Shapley | | | |
|---|---|---|---|---|
| **C** | FULL | SUBJECT | OBJECT | INTERCEPT |
| MM | 100 | 100 | 100 | 100 |
| MF | 94.6 | 100 | 89.4 | 100 |
| FM | 88.8 | 81.6 | 100 | 100 |
| FF | 84.6 | 83.0 | 92.2 | 100 |

(c) *%he>she*, stereotypical referents

| | GCD – **full** Shapley | | | |
|---|---|---|---|---|
| **C** | FULL | SUBJECT | OBJECT | INTERCEPT* |
| MM | 100 | 100 (100) | 100 (100) | 100 (88.0) |
| MF | 94.6 | 100 (99.6) | 95.4 (84.0) | 100 (84.8) |
| FM | 88.8 | 90.6 (77.4) | 100 (100) | 100 (91.0) |
| FF | 84.6 | 92.8 (75.6) | 97.4 (84.0) | 100 (89.2) |

(d) *%he>she*, stereotypical referents

Table 4: Results of Table 2, for both Shapley computations. Similar to Figure 5, it can be seen that the pattern remains the same, with the full Shapley computation again highlighting a slightly stronger male bias.