Appendix B: Questionnaire qualitative codes

The second author developed the following codes via a bottom-up process of interpreting common sentiments in listeners' responses to the final two questionnaire items: "How did the clips sound to you?" and "Do you have any other comments on the clips or on the survey?" The 79 listeners' responses were coded via these eight true-or-false codes. Below, we provide the codes, the number of responses that received these codes (with the total adding to greater than 79 since responses could receive multiple codes), the coding criteria, and examples.

DetectManip (16 TRUE): Coded TRUE if the listener indicated that they detected a pitch/voice quality manipulation (i.e., more than just excerpting or splicing).
- "A couple sounded robotic, like they had been modified."

QualityIssues (19 TRUE): Coded TRUE if the listener reported something wrong with the quality of the clips. (By contrast, numerous listeners commented positively on clip quality.) This was regardless of whether they detected a manipulation, though there was certainly overlap with DetectManip.
- "While the words were usually clear, the fact that many of the audio files had been altered was clear. On some of the samples, I wish there had been a third option: white, black, or *robot*"

CreakIssues (21 TRUE): Coded TRUE if the listener reported something wrong with the creaky phrases, explicitly or implicitly; substantial overlap with QualityIssues.
- "sometimes weird at the end"

TooShort (22 TRUE): Coded TRUE if the listener reported that the clips were too short to judge and/or seemed pragmatically unnatural as snippets of speech.
- "some sounded odd, very clipped and abrupt"
- "out of context"

RespStrategy (8 TRUE): Coded TRUE if the listener reported some sort of response strategy.
- "Obama's voice is so recognizable and I think of him as a Black man, so in a way, he always sounded Black to me. So I felt that I was rating him on an "Obama-specific" scale— like did he sound White for Obama? Or Black for Obama? Also I found myself trying to remember how I had answered the repeat clips previously, in order to be a nice reliable respondent, rather than answering based on my impression in that moment."

HardToJudge (23 TRUE): Coded TRUE if the listener reported difficulty making judgments on some or all trials; some overlap with TooShort.
- "I thought it was really hard to choose a lot of the time."

ObamaIsBlack (13 TRUE): Coded TRUE if the listener indicated that knowing that Obama is phenotypically black made it difficult to make "black-sounding"/"white-sounding" judgments; some overlap with HardToJudge.
- "If the goal is to try to discern the speaker's race by just the words/intonation used, maybe Obama isn't the best choice. Obviously his voice and face are well-known, so usually when I hear his voice I make the instant connection to his face—so to me he sounds black because I know he IS black. That said, given the followup questions, I'm guessing the voice needed to be Obama."

OtherDesignIssues (13 TRUE): Coded TRUE if the listener complained about some issue with the survey design and/or suggested changes to the design, aside from the quality of the clips and the difficulty of making judgments.
- "I really struggled to get through this survey. I wanted to mark every one of them black, since that's how Obama identifies. I had to take a few breaks."