We thank all the reviewers for their reviews! We will address the excellent writing/presentation-related suggestions in revision. Here we focus on clarifying questions about the framework, surveys, and results.

Re R1: Relevance and significance of results. We believe that the paper is appropriate for NeurIPS because it speaks to the applications of AI. Notably, we find that most people prefer machine-in-the-loop designs, and trust (except interpretability) is highly correlated with human preferences of delegability. We believe that working towards an improved and fine-grained understanding of public perception of AI on different tasks (both now and in deltas as the technology grows), will be valuable for researchers, industry leaders, and policy makers alike.

Re R1 & R2: Pearson r & p-value calculations in Table 2. Correlations in Table 2 are calculated by aggregating over individual likert ratings of the given component and the delegability level, for each individual response. This is done separately for the *personal* and *expert* surveys, resulting in a total of 28 tests. We did not apply Bonferroni correction because the dataset is relatively small and we chose to present the original results as an exploratory demonstration. After applying Bonferroni correction (14 tests each in the *personal* and *expert* surveys $= 28$ tests), most of the * and **s become non-significant. We will clarify this correction and follow R1's suggestion to include non-significant components in the appendix. Note that our empirical analysis in the main paper is limited by space constraints.

Re R1: Effect size. Our most significant findings have substantial effect size: 73.3% subjects prefer mixed-control, 18.4% prefer human only, and only 8.3% prefer complete automation; trust is the factor most correlated with delegability ($\rho \sim 0.5$). While we do discuss how the other effects are weaker than we expected in section 4.2, we could add some concluding discussion as well. The lack of clear effects also speaks to the challenge and lack of existing literature around this problem: we hope future studies can build on these results to help explain delegability more completely.

Re R2: Subjects' assumptions about AI/builders. Peoples' evaluations of trust undoubtedly depend on unmeasured assumptions, like who built the AI and why. Our goal is to advance our understanding of people's preferences towards delegation to AI, implicit assumptions included, because these same hidden assumptions are also present in current discussions about AI. While enumerating assumptions such as accuracy, financial costs, and builders of AI may lead to interesting results (given that we observe a significant correlation of value alignment), we leave that to future work.

Re R2: Conflation of the machine ability question with difficulty/delegability. First, we believe that human evaluation of trust in machine ability *does* include an implicit evaluation of difficulty for machines by definition, as it estimates the machine's ability to complete the task. Second, trust, difficulty, and delegability are all nebulous concepts; our work attempts to break down specific aspects of them. Third, machine ability does not fully explain delegability preferences (~ 0.5 in Table 2; lower than we would expect if people are conflating the two questions), and is at most weakly correlated with difficulty questions (see Figure 1 in Supplementary). For example, for "Reading bedtime stories to your child" in Table 3, we see reasonable trust in machine abilities (3.2/5), yet low delegability (1.8/4).

Re R2: Missing time/effort component. This was captured by the Effort component within Difficulty: "This task requires a great deal of time or effort to complete."

Re R2: Ask actual experts. Our primary interest is in understanding public preferences to delegate, so we administrated our surveys on Mechanical Turk, which does not typically involve experts. Also, given our diverse tasks, this would require experts for each task. That said, comparing expert-public differences is a great future direction.

Re R2: Knowledge of AI. We did not survey AI knowledge explicitly, but asked about computer knowledge. Subjects mostly rate themselves average/above, and there was no significant difference between them in delegability preferences.

Re R3: Choice of tasks. We source our task set from papers in AI conferences, daily life, occupations, and media coverage of AI. For example, "Analyzing and critiquing aesthetic qualities of photographs or other forms of art" comes from "Aesthetic Critiques Generation for Photos", ICCV 2017. These 100 tasks are meant as a first study; ideally, we build a large reference set that covers the entire "task space". Here the chosen set constitutes a reasonable starting point.

Re R3: Randomization of survey questions measuring components, order of delegability question, and survey procedure. The questions were not randomized, and the delegation question was always asked last. Each subject evaluated only one task. This ordering likely forced the subject to consider aspects of task delegability which they otherwise might not have, which could influence the decision, hopefully adding some depth. Because of the limited number of subjects (5 per task), question randomization may have introduced too much variability.

Re R3: Notion of coherence. By factor coherence, we mean the extent to which components within the factor may share information, as indicated by average pairwise correlations. High coherence justifies talking about the factors at a higher level, while low coherence indicates that more focus on individual components is warranted.

Re R3: High correlation between Value Alignment (VA) and Machine Ability (MA) in Trust. A nice future direction and we will add a discussion! Could be: evaluations of MA implicitly include VA; or VA is not considered an issue for tasks which currently have high MA; or an unknown component (e.g., builders of the AI from R2).