Supplementary Material: Artificial Agents Facilitate Human Cooperation through Indirect Reciprocity

Alexandre S. Pires[a,*] **and Fernando P. Santos**[a]

[a]University of Amsterdam, Amsterdam, The Netherlands

1 Cooperation index without artificial agents

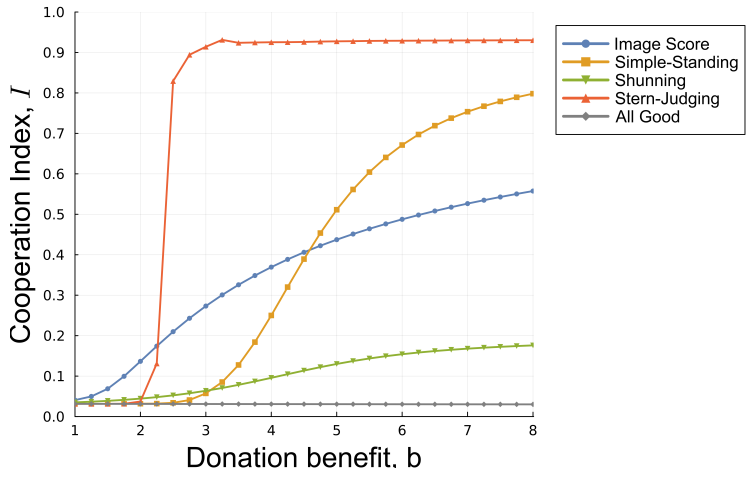

Figure 1. Cooperation index as a function of b, the donation benefit, when no AAs are present. We observe that changes in b result in different cooperation indexes across social norms. While all norms have low cooperation with a low b, cooperation quickly rises in **SJ**. A more gradual increase is present in **IS** and **SS**. In **SH**, cooperation levels remain low throughout the range. Finally, no cooperation is observed with **AG**, illustrating the impact of **IR**. The remaining parameters are identical to those of Figure 2 of the main text.

2 Cooperation index for different cost-benefit ratios

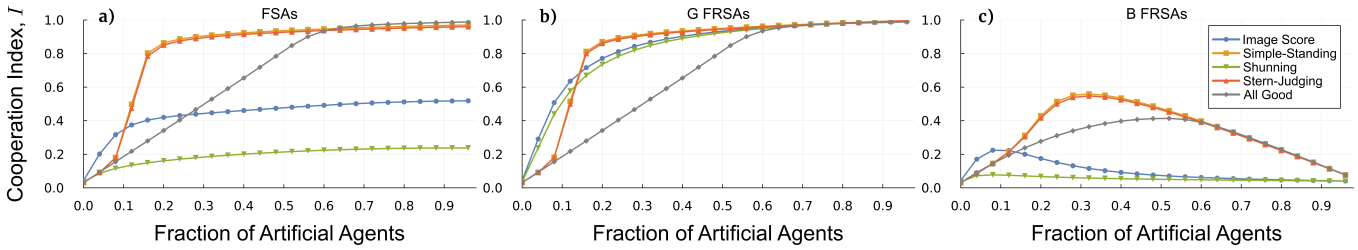

Figure 2. Cooperation levels for different fractions $\frac{A}{Z}$ of Discriminator AAs present in the population, for each social norm. The notation and parameters from Figure 2 of the main text are used, except with $b = 1.1$. **a)** FSAs (dynamic reputations). **b)** FRSAs are always perceived as good. **c)** FRSAs are always perceived as bad. Compared to the baseline with $b = 2$, **SS** and **SJ** require a higher amount of FSAs to reach peak cooperation. Despite the harder cooperation problem, **G** FRSAs still reach high cooperation values for **IS** and **SH**, even outperforming the other norms when in low prevalence. Bad FRSAs see a lower influence for **IS** and a later and softer peak for **SS** and **SJ**.

* Corresponding Author. Email: a.m.dasilvapires@uva.nl

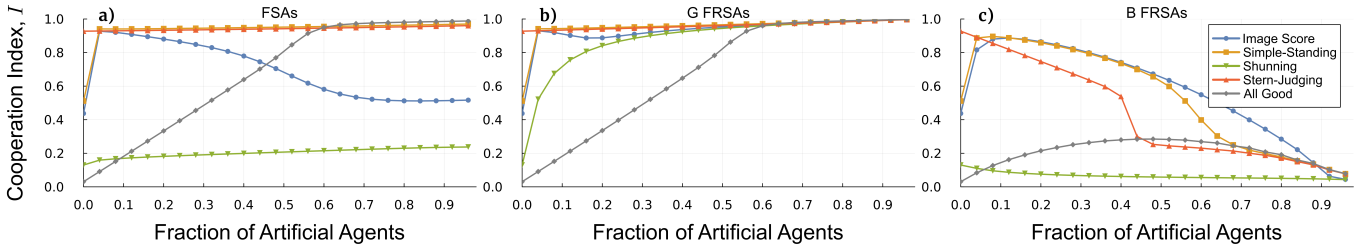

Figure 3. Cooperation levels for different fractions $\frac{\mathcal{A}}{Z}$ of Discriminator AAs present in the population, for each social norm. The notation and parameters from Figure 2 of the main text are used, except with $b = 5$. **a)** FSAs (dynamic reputations). **b)** FRSAs always perceived as good. **c)** FRSAs always perceived as bad. Compared to the baseline with $b = 2$, we observe considerably higher values of cooperation for **IS**, **SS** and **SJ** before the introduction of AAs. Although these values remain high for FSAs and **G** FRSAs under **SJ** and **SS**, with a slight drop in **IS** as more AAs are added, **B** FRSAs see a concrete decrease in cooperation.

3 Simplex under shunning with and without artificial agents

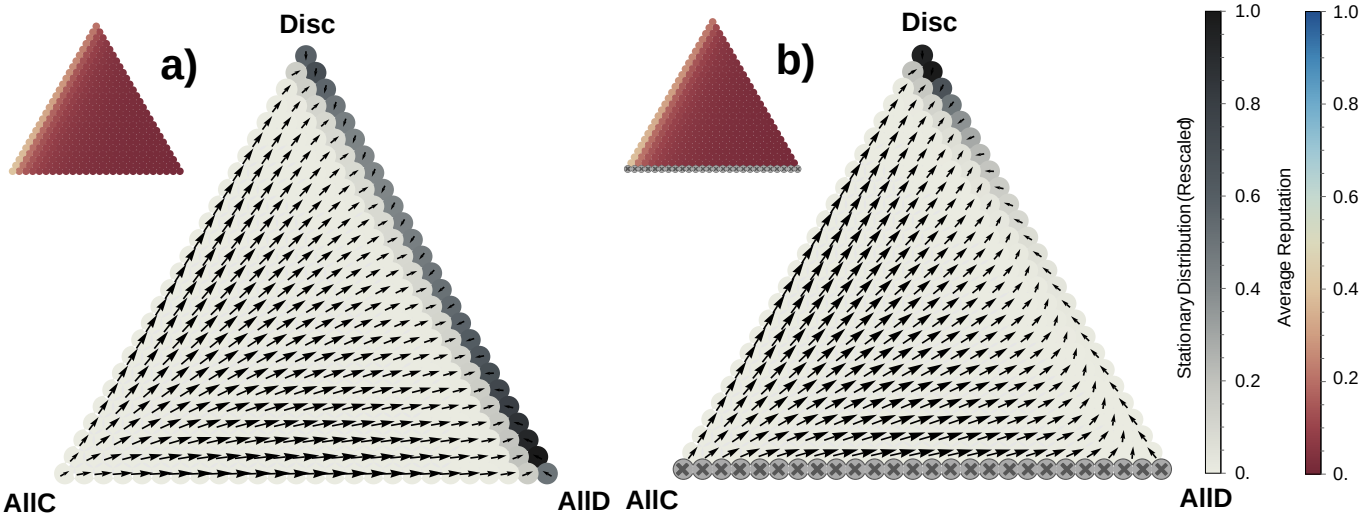

Figure 4. Evolutionary dynamics under **SH**. **a)** No AAs present in the population. **b)** Dynamic-reputation *Disc* FSAs are present in the population ($\mathcal{A} = 4$). The notation and parameters from Figure 3 of the main text are used. Without AAs, all the states in the *AllD-Disc* edge are common, however, since reputations are low, cooperation is rare. With AAs, *Disc* becomes the norm, however, cooperation is still low due to the average reputations being bad.

4 Gradient of selection for AllD-Disc simplex edge

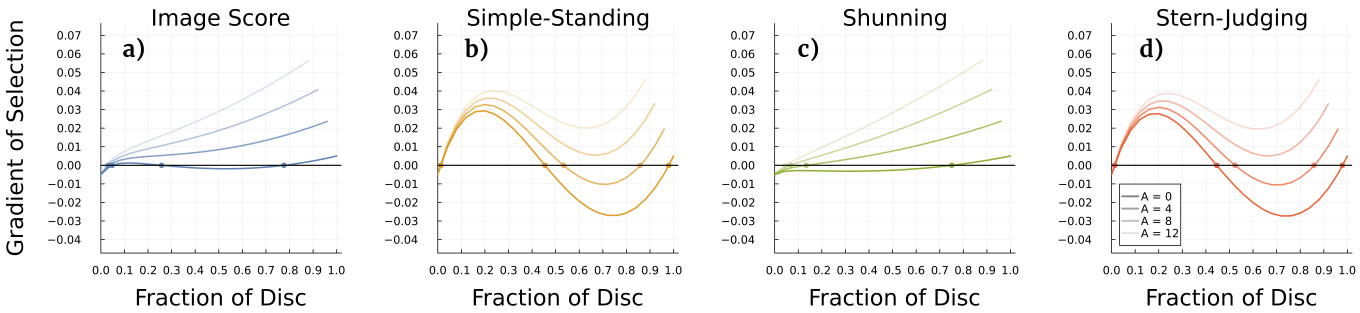

Figure 5. Gradient of selection on the *AllD-Disc* edge of the simplex ($n_{AllC} = 0$) under **a) IS**, **b) SS**, **c) SH**, and **d) SJ**. The X axis is identical to the *AllD-Disc* edge in the simplex, where a higher fraction of *Disc* means a lower fraction of *AllD* and vice versa. Whenever a line is above the X axis, the gradient of selection favors strategy *Disc*, otherwise it favors *AllD*. Lower opacity lines represent settings with a higher number of dynamic-reputation FSAs. We observe that **IS** and **SH** behave similarly, with the gradient close to neutral when AAs are not present, and tending more towards *Disc* the more AAs are added. **SS** and **SJ** also behave similarly, with a stable point of coexistence between the two strategies, and a translation along the Y axis as AAs are added. We note that for a given number of AAs, the gradient always favors *Disc*. The parameters used are identical to those of Figure 2 of the main text.

5 Dynamics without artificial agent imitation

Additional experiments were conducted by altering the strategy state transition probability function, given in Equation (13) of the main text, to consider the scenario where the adaptive agents do not imitate artificial agents. Under this scenario, only adaptive agents are considered as imitation role models, as AAs impact solely the average payoff of each strategy and limit the possible strategy and reputation states. As such, the transition probability will instead be given by

$$M'_{p\to p'}(n_{ijk}) = O^p_f(n_{ijk})\Big((1-\gamma)\frac{n_p}{Z}\frac{n_{p'} - T_f(p')}{Z-\mathcal{A}-1}P_{p\to p'}(n_{ijk}) + \gamma\frac{n_p}{2Z}\Big), \tag{1}$$

where, we recall, $T_f(p) = \mathcal{A}$ if $f_p = p$, and 0 otherwise.

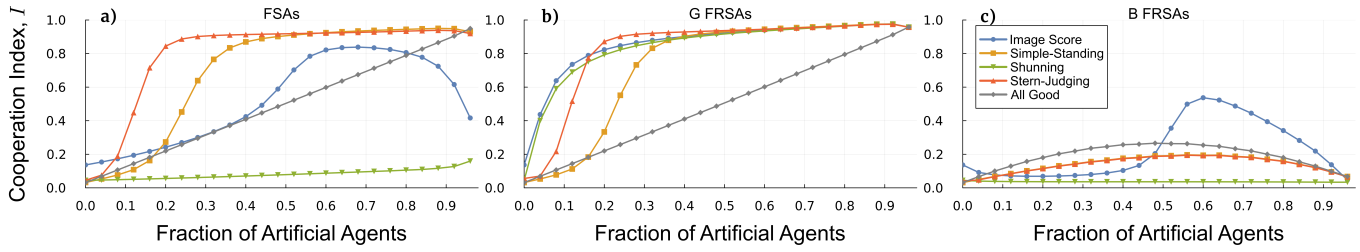

Figure 6. Cooperation levels for different fractions $\frac{\mathcal{A}}{Z}$ of Discriminator AAs present in the population, for each social norm, without allowing adaptive agents to imitate artificial agents. The notation and parameters from Figure 2 of the main text are used. **a)** FSAs (dynamic reputations). **b)** FRSAs always perceived as good. **c)** FRSAs always perceived as bad. Compared to the baseline with $b = 2$, we notice only a slight reduction in the impact of FSAs and **B** FRSAs when introduced in low numbers. We note a large difference for **B** FRSAs under **SS** and **SJ**, which now present a considerably lower efficacy, no longer resulting in an initial bump of cooperation.

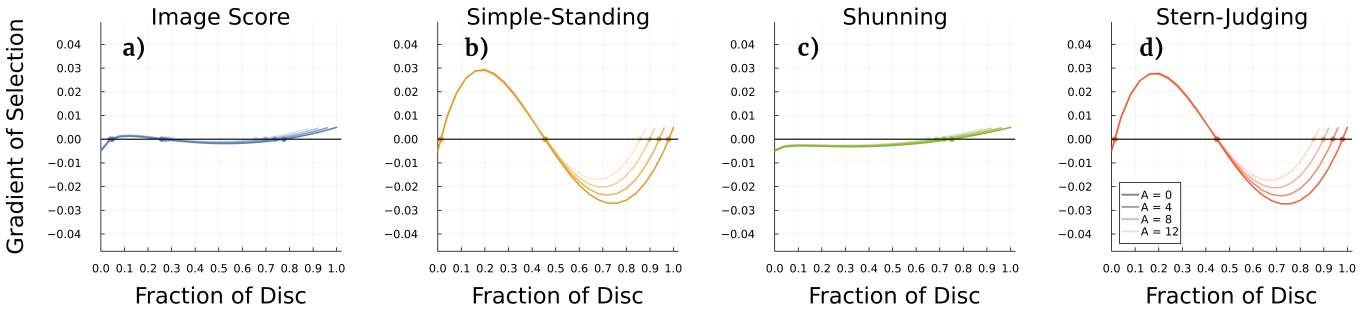

Figure 7. Gradient of selection on the *AllD-Disc* edge of the simplex ($n_{AllC} = 0$) when adaptive agents no longer imitate AAs, under **a) IS**, **b) SS**, **c) SH**, and **d) SJ**. We follow the notation of Figure 5. We again observe that the pairs **IS** and **SH**, and **SS** and **SJ** behave similarly. In contrast to the previous experiment, the tested number of AAs are not enough to favor *Disc* over *AllD*. However, in **IS**, the gradient is close to neutral, and in **SS** and **SJ** a weaker gradient towards *AllD* is observed as more AAs are added, suggesting that their effect is weakened but still existing. The parameters used are identical to those of Figure 2 of the main text.

6 Cooperation index with AllD and AllC artificial agents

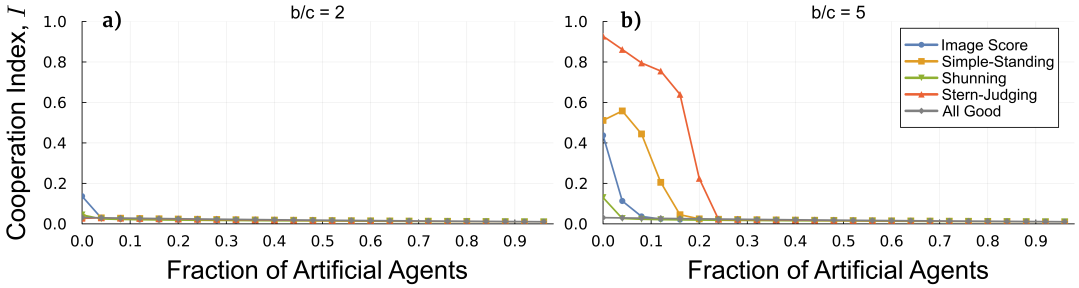

Figure 8. Cooperation levels for different fractions $\frac{A}{Z}$ of *AllD* agents present in the population, for each social norm. The notation and parameters from Figure 2 of the main text are used. **a)** $b = 2$. **b)** $b = 5$. Cooperation collapses quickly whenever *AllD* AAs are present, however, with a larger benefit b, a greater number of defector AAs is necessary to fully undermine cooperation.

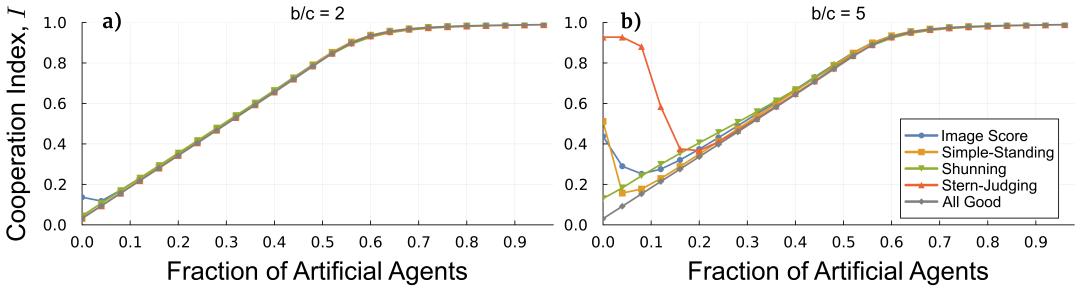

Figure 9. Cooperation levels for different fractions $\frac{A}{Z}$ of *AllC* agents present in the population, for each social norm. The notation and parameters from Figure 2 of the main text are used. **a)** $b = 2$. **b)** $b = 5$. We observe a correlation between the fraction of AAs added and the cooperation index, suggesting that *AllC* agents fail to promote cooperative behavior in the adaptive population, resulting instead in their exploitation. Furthermore, we observe that their presence can also reduce cooperation by motivating and sustaining exploitative behaviors.

7 Error study

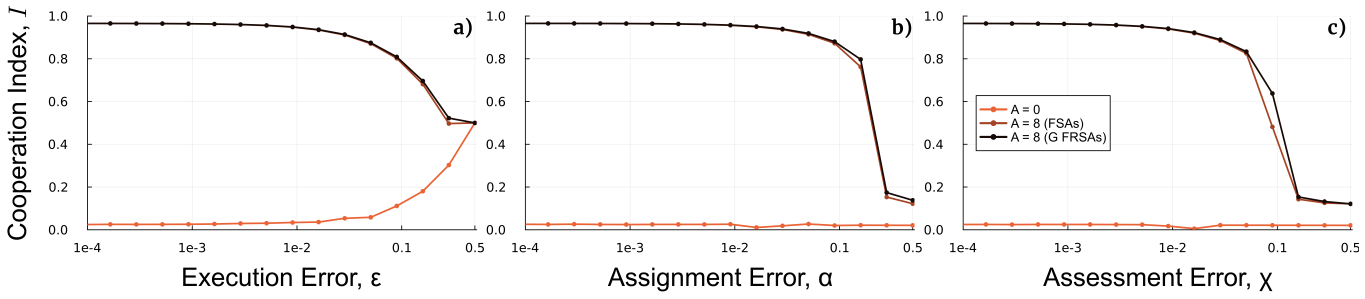

Figure 10. Cooperation index for different values of errors, under **SJ**. The lighter line corresponds to a scenario without FSAs, the darker line an environment with 8 dynamic-reputation FSAs, and the black line to a scenario with 8 **G** FRSAs. **a)** Execution Error (ϵ). **b)** Assignment Error (α). **c)** Assessment Error (χ). We observe that, if no FSAs are present, the cooperation index increases as the execution error increases. However, if artificial agents are present, the execution error will instead reduce cooperation. Varying the assignment and assessment errors shows little influence in cooperation without FSAs, possibly because the cooperation is already very low. With FSAs, we note a sharp drop in cooperation as the error rises above 0.15. While varying each error, the remaining types of errors are set to 0. All other parameters are identical to Figure 2 of the main text.

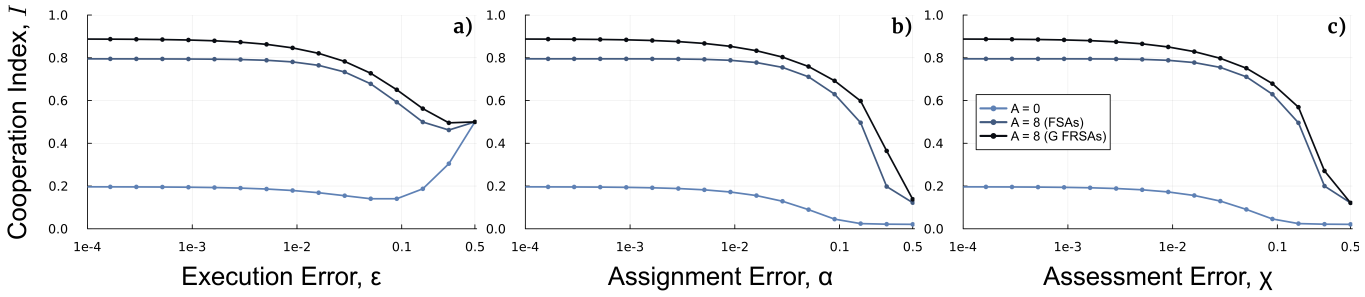

Figure 11. Cooperation index for different values of errors, under **IS**. We follow the same notation and parameters as Figure 10. Similar to Figure 10, if no FSAs are present, the cooperation index increases as the execution error increases. However, if artificial agents are present, the execution error will instead reduce cooperation. Higher assignment and assessment errors cause a lower cooperation, particularly with FSAs for errors above 0.10.

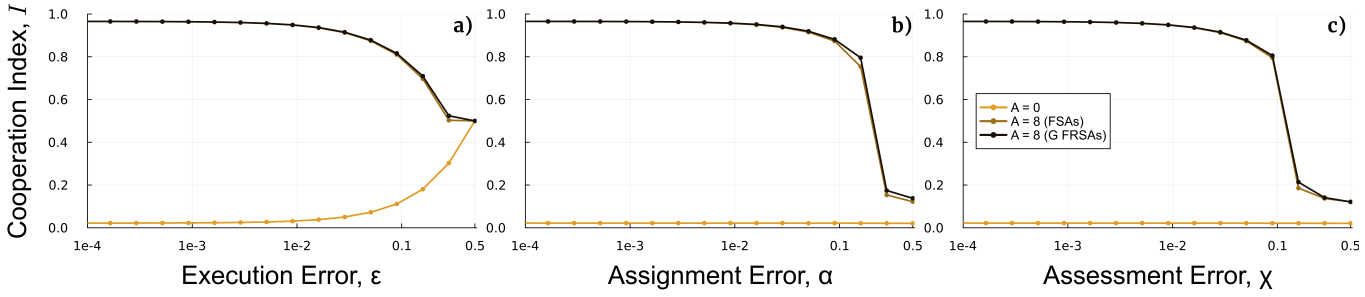

Figure 12. Cooperation index for different values of errors, under **SS**. We follow the same notation and parameters as Figure 10. We observe largely the same behavior, and therefore the same conclusions, as with **SJ**.

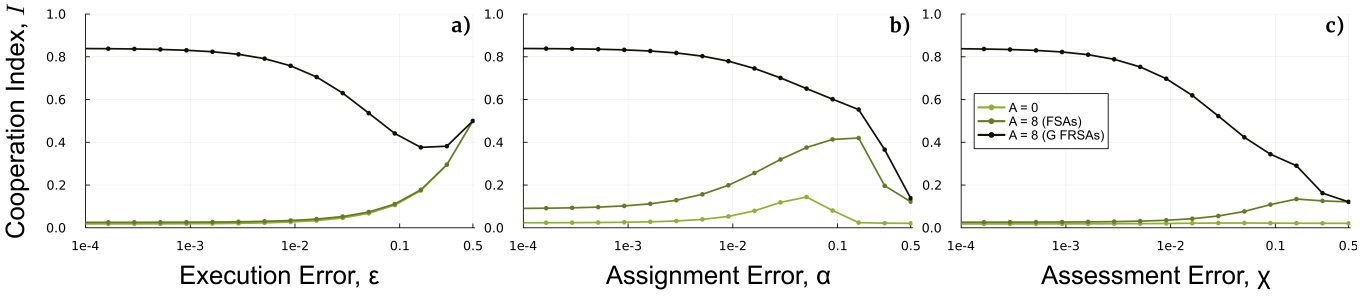

Figure 13. Cooperation index for different values of errors, under **SH**. We follow the same notation and parameters as Figure 10. We observe that the effect of errors is not as consistent as the remaining norms, with increases in the assignment and execution errors resulting in both increases and decreases in cooperation for the same scenario. Overall, errors will undermine cooperation with **G** FRSAs, while benefiting the remaining settings until a critical value for assignment and assessment errors, after which cooperation is again reduced.