Appendix: What if Alice is uncertain whether she and Bob are decision-entangled?[1]

Companion doc of Conditions for Superrationality-motivated Cooperation in a one-shot Prisoner's Dilemma.

(We'll assume Alice is 100% superrational (say she follows EDT and has no decision-theoretic uncertainty.)

Well, let's do some math.[2] We'll go with general payoffs this time.

Alice/Bob	C	D
C	a, a	b, c
D	c, b	d, d

Where $c > a > d > b$ and $b + c < 2a$.

How can we compute Alice's dominant strategy? Well, intuitively:

$U(C)_{Alice}$ = ($U(C)_{Alice}$ | [Bob cooperates]) · [probability that Bob cooperates given that Alice cooperated] + ($U(C)_{Alice}$ | [Bob defects]) · [probability that Bob defects given that Alice cooperated]

$U(D)_{Alice}$ = ($U(D)_{Alice}$ | [Bob cooperates]) · [probability that Bob cooperates given that Alice defected] + ($U(D)_{Alice}$ | [Bob defects]) · [probability that Bob defects given that Alice defected]

Now, let's try to replace text by mathematical terms :

- $U(C)_{Alice}$ | [Bob cooperates] = a
- probability that Bob cooperates given that Alice cooperated
  = probability of entanglement + (probability that Bob cooperates given no entanglement · probability of no entanglement)
  = $\alpha_{Alice} + \Gamma_{Alice}(1 - \alpha_{Alice})$

[1] Or what if she believes there is some form of ***partial decision-entanglement***? (My entanglement realist framework seems to suggest that entanglement regarding one unique decision can't be partial, but other frameworks might allow for this possibility.) I'm still assuming entanglement on one decision to be something binary, but a non-extreme credence in decision-entanglement is mathematically equivalent to a belief in some partial form of it, so it seems like there's no need to complicate things by making explicit the possibility of partial decision-entanglement. We can basically ignore it.

[2] My approach builds up on important things Nicolas Macé and Sylvester Kollin suggested in some informal discussion around superrationality. Cheers to them.

- $U(C)_{Alice}$ | [Bob defects] $= b$
- probability that Bob defects given that Alice cooperated
  = probability that Bob defects given no entanglement · probability of no entanglement
  $= (1 - \Gamma_{Alice})(1 - \alpha_{Alice})$
- $U(D)_{Alice}$ | [Bob cooperates] $= c$
- probability that Bob cooperates given that Alice defected
  = probability that Bob cooperates given no entanglement · probability of no entanglement
  $= (\Gamma_{Alice})(1 - \alpha_{Alice})$
- $U(D)_{Alice}$ | [Bob defects] $= d$
- probability that Bob defects given that Alice defected
  = probability of entanglement + (probability that Bob defects given no entanglement · probability of no entanglement)
  $= \alpha_{Alice} + (1 - \Gamma_{Alice})(1 - \alpha_{Alice})$

(α_{Alice} = probability Alice assigns to Bob and her being decision-entangled.)
(Γ_{Alice} = probability Alice assigns to Bob cooperating given no entanglement.)
($1 - \Gamma_{Alice}$ = probability Alice assigns to Bob defecting given no entanglement.)

We then get:
$U(C)_{Alice} = a(\alpha_{Alice} + \Gamma_{Alice}(1 - \alpha_{Alice})) + b((1 - \Gamma_{Alice})(1 - \alpha_{Alice}))$
$U(D)_{Alice} = c(\Gamma_{Alice}(1 - \alpha_{Alice})) + d(\alpha_{Alice} + (1 - \Gamma_{Alice})(1 - \alpha_{Alice}))$

Alice should cooperate iff $U(C)_{Alice} > U(D)_{Alice}$ ⇔ iff $\alpha_{Alice} > \frac{c\Gamma_{Alice} - d\Gamma_{Alice} + d - a\Gamma_{Alice} - b + b\Gamma_{Alice}}{a - a\Gamma_{Alice} - b + b\Gamma_{Alice} + c\Gamma_{Alice} - d\Gamma_{Alice}}$.

If we assume those specific payoffs,…

Alice/Bob	C	D
C	3,3	0,4
D	4,0	1,1

… that means that Alice should cooperate iff $\alpha_{Alice} > \frac{1}{3}$.

As you can see, we didn't use the framing with p *and* q as in the Normal PD and Perfect-copy PD, here, the main reason for that being that it was confusing and didn't work out when I tried. Computing $U(C)_{Alice}$ and $U(D)_{Alice}$ separately seems to work better, at least for my brain.

Now, if we assume those other payoffs, however,...

Alice/Bob	C	D
C	3,3	0,5

D	5,0	1,1

… that means that Alice should cooperate iff $\alpha_{Alice} > \frac{\Gamma_{Alice}+1}{\Gamma_{Alice}+3}$. Since $0 \leq \Gamma_{Alice} \leq 1$, we know here that Alice should cooperate iff α_{Alice} > [some number between $\frac{1}{3}$ and $\frac{1}{2}$, depending on the exact value of Γ_{Alice}].

Why does Γ_{Alice} matter in one case and not in the other?

Johannes Treutlein suggested that it may be because only the payoffs of the first PD (where $c = 4$) are *additively decomposable* (as defined by Oesterheld 2017, section 2.8.3). The next sub-section is an unpolished/informal proof that Γ_{Alice} never makes a difference in additively decomposable PDs. While this doesn't prove that Γ_{Alice}always matters in non-additively decomposable games (and I probably won't bother trying to prove that, since it doesn't seem helpful to my research project), that suggests that Johannes is likely to be right, I guess.

Anyway, it is interesting to notice that, depending on the exact payoffs of the PD, Alice may have to come up with some credence regarding the possibility that Bob cooperates given no entanglement (i.e., come with a value for Γ_{Alice}), which seems to make things more complicated. How she should/would do that is still an open question we could add to our earlier list.

Informal proof that Γ_Alice doesn't matter in additively decomposable games[3]

Let's assume that U_{Alice} and U_{Bob} are additively decomposable and symmetric. This means that:

$$U_{Alice}(a_{Alice}, a_{Bob}) = U_{\{Alice, Alice\}}(a_{Alice}) + U_{\{Alice, Bob\}}(a_{Bob})$$
$$U_{Bob}(a_{Alice}, a_{Bob}) = U_{\{Bob, Alice\}}(a_{Alice}) + U_{\{Bob, Bob\}}(a_{Bob})$$

Here, $U_{\{Bob, Alice\}}$ reads as "utility that Bob gets from Alice" and a_{Alice} as "Alice's action".

Now say that:
$U_{\{Bob, Alice\}}(C) = U_{\{Alice, Bob\}}(C) = z$
$U_{\{Bob, Alice\}}(D) = U_{\{Alice, Bob\}}(D) = f$
$U_{\{Alice, Alice\}}(C) = U_{\{Bob, Bob\}}(C) = g$

[3] Thanks a lot to Johannes Treutlein for his huge help regarding how to prove this. Mistakes are my own.

$U_{\{Alice,\ Alice\}}(D)$ = $U_{\{Bob,\ Bob\}}(D) = h$

Then, our payoff matrix is

Alice\Bob	C	D
C	g+z, g+z	g+f, h+z
D	h+z, g+f	h+f, h+f

So we have

- $a = g + z$
- $b = g + f$
- $c = h + z$
- $d = h + f$

This is the general set up of a PD with additively decomposable payoffs if I understood properly.

We know that Alice should cooperate iff $\alpha_{Alice} > \frac{c\Gamma_{Alice} - d\Gamma_{Alice} + d - a\Gamma_{Alice} - b + b\Gamma_{Alice}}{a - a\Gamma_{Alice} - b + b\Gamma_{Alice} + c\Gamma_{Alice} - d\Gamma_{Alice}}$.
Therefore, with our payoffs above, we get that Alice should cooperate iff
$\alpha_{Alice} > \frac{(z+h)\Gamma_{Alice} - (f+h)\Gamma_{Alice} + (f+h) - (z+g)\Gamma_{Alice} - (f+g) + (f+g)\Gamma_{Alice}}{(z+g) - (z+g)\Gamma_{Alice} - (f+g) + (f+g)\Gamma_{Alice} + (z+h)\Gamma_{Alice} - (f+h)\Gamma_{Alice}} \Leftrightarrow \alpha_{Alice} > \frac{h-g}{-f+z}$.

No Γ_{Alice} term! This means that Γ_{Alice} should be irrelevant to Alice's decision in PDs with additively decomposable payoffs.