			43	% agree	39,53%										
	b - can't just throw a niceness dataset and add corrigibility		17						assumptions						
number	name	section	do I agree	domination	no powerful helpers	unlimited opacity	betrayal is free	it is easier to go against humans than with	no bottlenecks in the environment	no instrumental abstraction convergence	need to solve in unbounded computation scenario	we NEED to get into the dangerous territory	comments		
39	only Eliezer Yudkowsky notices lethal difficulties	c overview of alignment as a field	?											0	
40	geniuses are non-transferable	c overview of alignment as a field	?											0	
41	this document is not enough to make someone a core alignment researcher. Eliezer Yudkowsky shouldn't be the only one to make it	c overview of alignment as a field	?											0	
42	there is no plan	c overview of alignment as a field	?											0	
43	this world does not look as if it's going to suirvive. people don't find enough flaws in their own plans. string theory was overrated.	c overview of alignment as a field	?											0	
10	does not generalize from safe to dangerous conditions	b.1 distributional leap	?						x					1	
12	drastic shift in distribution, new options	b.1 distributional leap	?						x					1	
35	AGIs will inevitably coalesce into a signle agent set against humanity	b.4 miscellaneous unworkable schemes	?						x					1	
5	cannot build a weak system, not enough	a lethal problem that needs to be solved and on the first try	?	x										1	
6	need to prevent other unaligned AGIs	a lethal problem that needs to be solved and on the first try	?	x										1	
7	no weak pivotal acts	a lethal problem that needs to be solved and on the first try	?	x										1	
36	we can't undestand its strategy	b.4 miscellaneous unworkable schemes	?		x								the action of finding a strategy is more costly than validating it	1	
11	it needs to generalize to build nanotechnology	b.1 distributional leap	no										STEM AI	0	
3	right on the first try	a lethal problem that needs to be solved and on the first try	no									x		1	
25	no idea what's inside matrices	b.3 central difficulties of sufficiently good and useful transparency/ interpretability	no							x				1	
28	we can't foresee all the options	b.3 central difficulties of sufficiently good and useful transparency/ interpretability	no		x									1	
30	no pivotal output that is humanly checkable	b.3 central difficulties of sufficiently good and useful transparency/ interpretability	no		x									1	
19	cannot point to noumena directly	b.2 central difficulties of outer and inner alignment	no							x			Paul Christiano abstractions graph	1	
17	inner alignment is unverifiable	b.2 central difficulties of outer and inner alignment	no			x	x							2	
2	independent bootstrapping of capabilities through any channel	a lethal problem that needs to be solved and on the first try	no				x		x					2	
4	there is a time limit and competing actors - need to go on	a lethal problem that needs to be solved and on the first try	no	x				x						2	
29	humans cannot inspect AGI output	b.3 central difficulties of sufficiently good and useful transparency/ interpretability	no		x	x								2	
26	knowing a medium strength is lethal won't help upgrade it and fix it, to compete with a potential stronger one	b.3 central difficulties of sufficiently good and useful transparency/ interpretability	no						x		x		stoppage would happen and the teams would talk it out	2	
20	human raters are bad	b.2 central difficulties of outer and inner alignment	no			x	x			x				3	
31	cannot verify behaviorally to see if it's deceitful	b.3 central difficulties of sufficiently good and useful transparency/ interpretability	no		x	x	x							3	
9	safe system would require maintenance, running a pivotal AGI is not passively safe	a lethal problem that needs to be solved and on the first try	yes											0	
1	alpha go self-learning faster than a human	a lethal problem that needs to be solved and on the first try	yes											0	
8	capability generalization across domains	a lethal problem that needs to be solved and on the first try	yes											0	
13	new alignment problems at superintelligence	b.1 distributional leap	yes											0	
14	only evaluates lethal options once sufficiently powerful	b.1 distributional leap	yes											0	
15	fast capability gains break alignment - humans as an example	b.1 distributional leap	yes											0	
16	outer optimizaiton not enough	b.2 central difficulties of outer and inner alignment	yes											0	
18	no Cartesian ground truth	b.2 central difficulties of outer and inner alignment	yes											0	
21	Capabilities generalize further than alignment once capabilities start to generalize far.	b.2 central difficulties of outer and inner alignment	yes											0	
22	core structure for reasoning emergent in low-entropy high-structure environments, no such for alignment	b.2 central difficulties of outer and inner alignment	yes											0	
23	corrigibility is unsolvable in the limit	b.2 central difficulties of outer and inner alignment	yes											0	
24	CEV right on first try vs corrigible yet not lethal	b.2 central difficulties of outer and inner alignment	yes											0	

			43	% agree	39,53%										
	b - can't just throw a niceness dataset and add corrigibility		17	assumptions											
number	name	section	do I agree	domination	no powerful helpers	unlimited opacity	betrayal is free	it is easier to go against humans than with	no bottlenecks in the environment	no instrumental abstraction convergence	need to solve in unbounded computation scenario	we NEED to get into the dangerous territory	comments		
27	optimizing for transparency optimized for unaligned undetected thoughts	b.3 central difficulties of sufficiently good and useful transparency/ interpretability	yes											0	
33	AI is alien and incomprehensible	b.3 central difficulties of sufficiently good and useful transparency/ interpretability	yes											0	
37	most people are business as usual sheeple	c overview of alignment as a field	yes											0	
38	AI safety is not remotely productive. only easy problems are approached, produces mostly noice	c overview of alignment as a field	yes											0	
32	human thoughts and abstractions are not AGI-capable	b.3 central difficulties of sufficiently good and useful transparency/ interpretability	yes						x				STEM AI	1	
34	humans can't join a conspiracy among AGIs. we can't reason about their code	b.4 miscellaneous unworkable schemes	no		x	x	x							3	
				4	6	5	5	1	6	3	1	1	5		