

Keyword Index

Accessibility	91
Adaptive design	8
Adversarial Attacks	112
agent incentives	22
agent type	82
AGI	44
AGI safety	44
AI for Safety	1
AI Safety	75, 130
AI safety	105
AI Safety in Healthcare	119
AI Safety Landscape	1
AI Saftey	8
AI value alignment	75
AIXI	105
algorithmic fairness	37
artificial intelligence	68
Artificial Intelligence	61
Augmented Utilitarianism	75
Autonomous Driving	61
Autonomous Systems	57
autonomous vehicles	82
Bayesian learning	105
Behavioral Safety	61
Bias	91
Black-Box Attacks	112
causal graphs	105
causality	44
Clinical trial	8
combinatorial testing using ontologies	123
Commitments	98
conservative	29
Coordination and Cooperation	98
corrigibility	29
counts as	68
crowdsourcing	82
deep learning	37, 51
Deep Neural Networks	57
Deep Q-learning from Demonstrations	112
Deep Reinforcement Learning	112, 137
Dependability	57

DL	37
Dynamic Safety Cases	1
embedded agency	105
Empathy	130
Ethical Goal Function	75
ethics	82
Explainable and Safe AI	119
frameworks	44
Healthcare	119
impact	29
impact measures	22
influence diagrams	44
inverse reinforcement learning	37
IRL	37
learning	37
Learning from Demonstrations	112
Machine Ethics	130
machine learning	15, 37, 51
markov decision processes	37
Markov Decisions Processes	98
MDPs	37
misalignment	105
ML	37
modal logic	68
moral dilemmas	82
multi-armed bandit	8
Multi-Task Learning	137
objective specification	22
Perception	57
perturbations	51
Planning under Uncertainty	98
Policy Imitation	112
preferences	15
Reinforcement Learning	61, 130
reinforcement learning	22, 29, 37, 105
review	44
RL	37
Safety	57

Safety of Autonomy	1
Safety-Criticality	61
Sequential Triggers	137
side effects	22, 29
Software engineers	91
system testing for autonomous systems	123
taxonomy	105
test automation	123
Topological data analysis	91
transparency	82
trolley problem	82
trust	68
Uncertainty	57
uncertainty	37
value alignment	15
value learning	37
verification	68
virtual reality	82
Watermarking	137
wireheading	105