Adversarial goals	Adversarial knowledge	Exemplary adversarial capabilities	SEA AI attack effect
Epistemic spear phishing	On AI tools: black-box, grey-box or white-box setting On human victims: grey-box settings (e.g., via open-source intelligence gathering)	Present-day-AI-*only*-assembled messages (via e.g., deepfake text) featuring attention-mongering science news to lure scientists into clicking on malicious links with confirmatory contents	In deepfake era, alarmist science journalism risks to make scientists vulnerable to spear phishing worsening cybersecurity
Epistemic denial-of-service attack	On AI tools: black-box, grey-box or white-box setting On human victims: black-box or grey-box settings	Flood of present-day-AI-*only*-assembled submissions to scientific venues (including e.g. rehearsing old assumptions using authoritatively perceived deepfake papers and/or via deepfake experiments and/or spreading doubt about new selected target theory via authoritatively sounding deepfake papers and/or via deepfake experiments) leading to overwhelmed reviewers causing poorer-quality human peer-review, deepfake peer-review or poorer-quality hybrid peer-review	Peer-review, experiments risk to be outsourced to present-day AI, deepfake papers that by definition cannot contain better new yet unknown theories about the world risk to pollute the information ecosystem
Epistemic injection	On AI tools: black-box, grey-box or white-box setting On human victims: black-box or grey-box settings	Insertion of misguiding, convincingly sounding present-day-AI-*only*-assembled information into scientific papers (e.g., via copy and paste of deepfake text) to mislead or slow down science via e.g., fear narratives on AI itself	Risk of epistemic stagnation of science
Epistemic spyware	On AI tools: black-box, grey-box or white-box setting On human victims: black-box or grey-box settings	Perform epistemic injection with the purpose of tracing inattentive scientists or scientists with a comparatively weaker epistemology via naïve citations	Scientists with weaker epistemologies risk to be ridiculed in the future
Epistemic scareware	On AI tools: black-box, grey-box or white-box setting On human victims: grey-box settings (e.g., via open-source intelligence gathering)	Build closed-source generative AI with personalized backdoor leading to a remotely human-assisted mode in the background (which the victim does not know) such that vulnerable scientists afraid of artificial superintelligence become convinced that system is superintelligent, get distressed and interpret non-reproducibility as intentional malicious deception of present-day AI itself.	Risk for scientists to develop mental health issues including delusions concerning a harmful omnipotent artificial quality superintelligence

Epistemic spoofing	On AI tools: black-box, grey-box or white-box setting On human victims: black-box or grey-box settings	Build up deepfake persona camouflaged as deepfake replica of existing scientists on social media using e.g., deepfake images of actually existing scientists, generate deepfake experiments, deepfake papers, deepfake peer review to further a specific agenda	Scientists with weaker epistemologies risk to be impersonated more plausibly – weakening the perceived value of science in a society
Epistemic rootkit	On AI tools: black-box, grey-box or white-box setting On human victims: grey-box settings (e.g., via open-source intelligence gathering)	Build up deepfake persona on social media as simulacrum of never existing scientists of recognized universities to infiltrate scientific peer review and automatically gain privileges via e.g., deepfake peer review to further a specific agenda	Risk of structurally supported epistemic distortion in science
Epistemic virus	On AI tools: black-box, grey-box or white-box setting On human victims: black-box or grey-box settings	Perform targeted epistemic injection using deepfake text and/or deepfake experiments with a misguided idea such as quality artificial superintelligence to cause a worldwide epistemic distortion with many followers	Existential risk through delusional scientists that unintentionally fuel catastrophic reactions of malicious people
Epistemic zero-day	On AI tools: black-box, grey-box or white-box setting On human victims: grey-box settings (e.g., via open-source intelligence gathering)	This is a generic ability. Here only one thinkable example: automatize microtargeted deepfake paper assembly tailored to specific reviewers to drastically increase acceptance of misleading contents.	Pleasant deepfake bubbles risk to slow down advance of science
Epistemic Trojan horse	On AI tools: black-box, grey-box or white-box setting On human victims: black-box or grey-box settings	Make the quality superintelligence narrative appear as scientific via submitting authoritatively sounding deepfake papers citing popular AI researchers, later write more papers citing the Trojan horse as proof for need of AI pause	Scientists of Western civilization with weaker epistemologies can be epistemically fooled by adversaries that know better
Epistemic supply chain attack	On AI tools: black-box, grey-box or white-box setting On human victims: black-box or grey-box settings	Pollute internet with deepfake experimental material to further a specific agenda or in general destabilize empiricist science via domino-effects	Risk of epistemic self-sabotage of science relying on data
Epistemic man-in-the-middle attack	On AI tools: black-box, grey-box or white-box setting On human victims: black-box or grey-box settings	Steer deepfake summaries via “summarizing” language models to alter perception of contents of scientific information and subtly insert own agenda	Epistemic derailing via fragmented incoherent observation statements

Epistemic buffer overflow	On AI tools: black-box, grey-box or white-box setting On human victims: black-box or grey-box settings	Interfere with new better short-term memory formation of scientists to avoid long-term memory consolidation by generating deepfake flood of confirming science narratives that overwrite just acquired short-term memories	Risk of suppression of innovations
Epistemic watering hole attack	On AI tools: black-box, grey-box or white-box setting On human victims: black-box or grey-box settings	Spread epistemic Trojan horses in renowned journals that are regularly read by most scientists	Risk of undermined scientific institutions losing credibility
Epistemic astroturfing	On AI tools: black-box, grey-box or white-box setting On human victims: black-box or grey-box settings	Generate deepfake flood of information to confirm scientific misinformation narratives on social media	Risk of world-wide disinformation via weaker epistemologies in science
Epistemic data-poisoning	On AI tools: black-box, grey-box or white-box setting On human victims: black-box or grey-box settings	Before conducting adversaries' desired study, flood internet with deepfake data that confirms the study	Risk of disconnect and alienation from real-world problems negatively impacting grip on the world leading to uncontrolled catastrophes
Epistemic session hijacking	On AI tools: black-box, grey-box or white-box setting On human victims: black-box or grey-box settings	Use multiple AI-driven avatars in scientific VR conferences impersonating same known scientist to convince people of quality superintelligence	Riks of psychological manipulation of scientists
Epistemic shortcut	On AI tools: black-box, grey-box or white-box setting On human victims: black-box or grey-box settings	As human crowd worker, use deepfake outputs for studies one has to submit as crowd worker – which introduces more deepfake experiments in science	Risk of inefficient misinformed policies with no fruitful results
Epistemic delusion	On AI tools: black-box, grey-box or white-box setting On human victims: black-box or grey-box settings	Let scientists assign agency and consciousness to present-day AI by secretly using AI-executed tricks from the psychology and neuroscience of magic making scientists believe that present-day itself is the magician	Risk of world dominance by other nations (including adversaries) with a stronger epistemology leading to wars
Epistemic babble	On AI tools: black-box, grey-box or white-box setting On human victims: black-box or grey-box settings	Create deepfake material corroborating empiricist post-truth narratives that the world became incomprehensible	Existential risk for humanity via second-order harm in connection with malicious actors

Epistemic stigmatization	On AI tools: black-box, grey-box or white-box setting On human victims: black-box or grey-box settings	Create convincing deepfake papers about success of deepfake detection letting human scientists believe that present-day AI can detect present-day AI	Risk of stigmatization, dehumanization and suppression of scientists that are human statistical outliers and whose research outputs are misclassified as AI-generated leading to decrease of cognitive diversity in science
Epistemic subordination	On AI tools: black-box, grey-box or white-box setting On human victims: grey-box settings (e.g., via open-source intelligence gathering)	Create micro-targeted personalized deepfake-text based information campaign to convince scientists of omnipotent AI narrative, leading to a loss of agency and a will to surrender epistemic control to global entities	Risk to democracy by exaggerated claims of scientists about present-day AI capabilities leading to unnecessary global enforcement of superfluous policies leading to international security risks
Epistemic paralysis	On AI tools: black-box, grey-box or white-box setting On human victims: grey-box settings (e.g., via open-source intelligence gathering)	Create micro-targeted personalized deepfake-text based scenario where scientist believes artificial quality superintelligence has been achieved via recursive self-improvement from narrow AI, make scientist refrain from working on novel theories out of fear and e.g., promise to only publish deepfake papers	Risk of personalized epistemic terror that is not directly apparent from the outside, suicide risks
Epistemic ransomware	On AI tools: black-box, grey-box or white-box setting On human victims: grey-box settings (e.g., via open-source intelligence gathering)	First create epistemic paralysis, then demand money to soothe the omnipotent AI	Risk of large financial losses by scientists