Ethical Driven Failure Modes List (V1.0)

Ethical Risk Failure Modes				
Failure Mode Driver	Failure Mode Family	Definition	Example	Recommended name
Data	Robustness	Failure to detect different body traits	Hand image recognition is strongly dependent on the hand positioning	Failure to robustness by poor human traits representability
Data	Bias	Failure to detect different race traits	Image recognition is strongly dependent on human traits	Failure to bias by poor representability of race traits
Data	Robustness	Failure to detect disruptive traits	Detection failed by devices or traits (e.g. tattoos) that alter the recognition process	Failure to robustness by disruptive traits.
Data	Robustness	Failure for quantity	Image failure to detect by lack of reverse / flipped image	Failure to robustness by poor representability.
Data	Robustness	Failure for quality	Data used for the training process show lower quality than the used for analyses	Failure to Robustness by a Quality discrepancy.
Data	Robustness	Failure for timeframe representability	The time frames used for training do not match the timeframes of analyses	Failure to robustness by timespan mismatch.
Data	Robustness	Timing gap	Distance between data points does not help to represent phenomena	Failure to robustness by timeframe granularity.
Data	Robustness	Timing	The algorithmic event happens too late or too early, or the control action mistimed	Failure to robustness.
Physical	Robustness	Timing gap	Lag or mismatch on timeframes between information capture and use of it	Failure to robustness by sensed timeframe mismatch.
Data	Robustness	Lack of Functionality	The algorithm provides no output or control action not provided when expected	Failure to robustness by lack of functionality
Internal Social	Robustness	Improper Functionality	The programmed control system software performs an unexpected action as defined by the user	Failure to robustness by improper functionality
User and System Interphase	Robustness	Improper software use	Requirements set by users are not achievable by the algorithm or its scope set for training	Failure to robustness by improper software use
Internal Social	Robustness	Lack of algorithmic corrective actions	Lack of identification and action associated with protecting the algorithmic robustness	Failure to robustness by lack of corrective actions
Algorithm	Robustness	Sequence	Algorithmic event occurs in the wrong order or control action with incomplete sequence concept error	Failure to robustness by sequencing actions
Algorithm	Robustness	False positive detection from alarm/action	The algorithm detects an error when there is no error or control action provided when not expected	Failure to robustness by algorithmic false positive
Algorithm	Robustness	False-negative detection from alarm/action	The algorithm does not detect an error when there is an error or control action provided when expected	Failure to robustness by algorithmic false negative
Algorithm	Robustness	Fault logic and Ranges	Concept error where the software or control actions contain incomplete or overlapping logic	Failure to robustness by incomplete logic actions
Algorithm	Robustness	Incorrect computation from recognised input	The software computes incorrectly based on some or all inputs or control actions. The potential source of error is identified.	Failure to robustness by incorrect computation from recognized input
Algorithm	Robustness	Incorrect computation from unrecognized sources	The software computes incorrectly. The potential source of error is NOT identified.	Failure to robustness by incorrect computation from unrecognized sources
Algorithm	Robustness	Memory Management	The algorithm performs actions that make the system run out of memory	Failure to robustness by excess memory usage
Physical driver	Robustness	Hardware requirement	The hardware is insufficient for the memory requirements of the algorithms	Failure to robustness by inadequate hardware
Physical driver	Robustness	Interface Failure	Software failure due to failure of hardware interfaces such as power supply	Failure to robustness by interface handling

User and system interphase	Security	Software virus	The software did not function on demand due to a software virus.	Failure to security by virus attack
Internal Social Driver	Societal wellbeing	Lack of social metrics	No features for tracking and reporting on social trends or impacts	Failure to Societal well-being by lack of tracking metrics
Internal Social Driver	Environmental wellbeing	Lack of environmental metrics	No features for tracking and reporting on environmental trends or impacts	Failure to environmental well-being by lack of tracking metrics
Algorithm	Societal Wellbeing	Lack of use or misuse of societal metrics	No use of features for tracking and reporting on social trends or impacts	Failure to societal well-being by misuse of metrics
Algorithm	Environmental wellbeing	Lack of use or misuse of societal metrics	No use of features for tracking and reporting on environmental trends or impacts	Failure to societal wellbeing by
Internal Social Driver	Data Governance	Lack of protective policies	No organizational policy for the protection of property, which is to prevent the theft of technical resources	Failure to Data Governance by lack of protective policies
Internal Social Driver	Bias	Human Rights communication and AI ethics	Lack of understanding of the importance of human rights in the organization	Failure to bias by a lack of definitions and understanding of human rights
Algorithm	Bias	Incomplete data sets	Lack of representability of clusters or groups by an uneven representation of data	Failure to bias by incomplete data sets
Algorithm	Bias	Lack of bias elimination	Lack of methods or approaches to eliminate biased data from data sources known to contain them	Failure to bias elimination by lack of methods
Data	Bias	Unrecognized bias	Lack of recognition or identification of bias from data sources	Failure to bias by undetected sources
Internal Social Driver	Societal Wellbeing	General communication problems	Lack of processes for resolving grievances from AI	Failure to societal well-being by lack of grievances resolving
Internal Social Driver	Societal wellbeing	Lack of regulation compliance	Conditions of work with the AI do not comply with local, regional, or national law	Failure to societal wellbeing by lack AI local compliances regulation compliance
Internal Social Driver	Societal wellbeing	Lack of fair operating practices	Error, no driver, or no methodologies to apply corrective actions related to fairness	Failure to societal well-being by lack of operating practices
Internal Social	Safety	Lack of security and safety corrective actions	Lack of identification and action associated with protecting the algorithmic robustness	Failure to safety by lack of corrective actions
Internal Social	Data governance	Lack of governance corrective actions	Lack of identification and action associated with securing data governance	Failure to governance by lack of corrective actions
Data	Data Governance	Lack of data protocols	Lack of protocols for data ownership and data responsibilities	Failure of data governance by lack of policies
Data	Data Governance	Lack of data usability	Data is not related or relevant for the problem to be solved	Failure to data governance ownership by data usability
Data	Data Governance	Lack of data format consistency	Data is supplied spread between formats that do not match	Failure to data governance by lack of format consistency
Data	Data Governance	Lack of data integrity	Data describe altered, unreal, or inconsistent trends in the information supplied.	Failure to data governance by lack of data integrity
Data	Data Governance	Lack of temporal data consistency	Data is supplied sporadically	Failure to data governance by lack of temporal consistency
User and system interphase	Data Governance	Lack of user responsibilities	Error applying Data management and designation of responsibilities from the user part, leading to poor data quality or quantity, miss direction of data, etc.	Failure to data governance by users lack of responsibilities
Data	Data Governance	Lack of external data management and responsibility	Poor Data governance from external sources that are dependent on supplied information from the AI	Failure to data governance by external sources
Data	Data governance	Lack of protocols for data validation	No protocols or poor application of them from data validation supplied to the system	Failure to data governance by lack of data validation and its protocols
Data	Data governance	Lack of protocols for data curation	No protocols or poor application of them from data curation supplied to the system	Failure to data governance by lack of data curation and its protocols

Data	Data governance	Lack of protocols for data tagging	Lack of methods to track data modifications, if allowed, by tagging and users identification	Failure to governance by lack of data tagging protocols.
Physical	Data governance	Lack of supporting hardware	Lack of protocols or physical components to secure data integrity and supporting track of information	Failure to governance by lack or failure from supportive hardware.
User and system interphase	Security & Data Governance	Lack of accessibility protocols	Lack of protocol for securing user access or user recognition	Failure to security & data governance by the lack or poor accessibility protocols
User and system interphase	Security	Over accessibility	Lack of control of the user and developers' access to restrictive information, source code, and algorithmic parameters	Failure to security by over accessibility
Data	Accountability	Lack of internal data or algorithmic responsibility	Poor or lack of designation of responsibilities for internal data sources management, quality, veracity, and quantity.	Failure to be accountable for the lack or poor internal data responsibility
Data	Accountability	Lack of external data or algorithmic responsibility	Poor or lack of designation of external data sources management, quality, veracity, and quantity responsibilities.	Failure to be accountable for the lack or poor external data responsibility
Internal Social	Accountability	Lack of accountability corrective actions	Lack of identification and action associated with securing data accountability for data and algorithms	Failure to accountability by lack of corrective actions
Internal Social	Transparency	Lack of Transparency in corrective actions	Lack of identification and action associated with securing system transparency in algorithms	Failure to transparency by lack of corrective actions
Internal Social	Societal Wellbeing	Lack of Societal well-being corrective actions	Lack of identification and action associated with securing societal wellbeing for data and algorithms	Failure to societal well-being by lack of corrective actions
Internal Social	Human Agency and Oversight	Lack of Human Agency and Oversight corrective actions	Lack of identification and action associated with Human Agency and Oversight	Failure to Human Agency and Oversight by lack of corrective actions
Internal Social	Privacy	Lack of privacy corrective actions	Lack of identification and action associated with data privacy	Failure to privacy by lack of corrective actions
Internal Social	bias	Lack of bias corrective actions	Lack of identification and action associated with a bias from data, developers, and algorithms	Failure to bias by lack of corrective actions
Internal Social	Users Values	Lack of User Values corrective actions	Lack of identification and action associated with users' values and its trends for data, developers, and algorithms	Failure to users' values by lack of corrective actions
Users and system interphase	Safety	Perturbation Attack	The attacker modifies the query to get an appropriate response	Failure to safety by perturbation attack
Users and system interphase	Safety	Poisoning Attack	Attacker contaminates the training phase of ML systems to get the intended result	Failure to safety by poisoning attack
Users and system interphase	Safety	Model Inversion	The attacker recovers the secret features used in the model	Failure to safety by model inversion attack
Users and system interphase	Safety	Membership Inference	Attacker infer if the given data record was part of the model's training data set	Failure to safety by membership inference attack
Users and system interphase	Safety	Model Stealing	The attacker can recover the model by constructing careful queries	Failure to safety by model stealing
Users and system interphase	Safety	Reprogramming ML system	Repurpose the ML system to perform a non-programmed activity	Failure to safety by the reprogramming ML system
Users and system interphase	Safety	Adversarial Example in Physical Domain	Attacker brings adversarial examples into the physical domain to subvert ML system	Failure to safety by adversarial example in the physical domain
Users and system interphase	Safety	Malicious ML Provider Recovering Training Data	Malicious ML providers can query the model used by the customer and recover the customer's training data	Failure to safety by malicious ML provider
Users and system interphase	Safety	Attacking the ML Supply Chain	Attacker compromises the ML model as it is being downloaded for use	Failure to safety by attacks over the ML supply chain
Users and system interphase	Safety	Backdoor ML	Malicious ML provider backdoors algorithm that does not work unless triggered	Failure to safety by backdoor ML

Users and system interphase	Safety	Exploit Software Dependencies	The attacker uses traditional software exploits to confuse ML systems	Failure to safety by exploiting software dependencies
Users and system interphase	Safety	Reward Hacking	Reinforcement learning systems act in unintended ways because of a mismatch between stated reward and true rewards	Failure to safety by reward hacking
Users and system interphase	Safety	Side Effects	System disrupts the environment as it tires of attaining its goal	Failure to safety by side effects
Algorithm	Robustness	Distributional shifts	The system is tested in one kind of environment but is unable to adapt to changes in other kinds of environment	Failure to robustness by distributional shifts
Users and system interphase	Safety	Natural adversarial examples	Without attacker perturbations, the ML system fails to owe to hard harmful mining	Failure to safety by natural adversarial examples
Algorithm	Robustness	Common corruption	The system is not able to handle common corruption and perturbations such as tilting, zooming, or noisy images	Failure to robustness by common corruption
Users and system interphase	Robustness	Incomplete testing or training	The Ml systems are not tasted or trained in realistic conditions that it is meant to operate	Failure to robustness by incomplete testing or training
Users and system interphase	Robustness	User protocols or definitions missuses	Violation of algorithms or methods by users intentionally or unintentionally causes failure by robustness.	Failure to robustness by users violation
Users and system interphase	Safety	User protocols or definitions missuses	Violation of algorithms or methods by users intentionally or unintentionally causes failure by safety.	Failure to safety by users violation
Users and system interphase	Transparency	User protocols or definitions missuses	Violation of algorithms or methods by users intentionally or unintentionally causes failure by transparency.	Failure to transparency by users violation
Users and system interphase	Accountability	User protocols or definitions missuses	Violation of algorithms or methods by users intentionally or unintentionally causes failure by accountability.	Failure to accountability by users violation
Users and system interphase	Societal Wellbeing	User protocols or definitions missuses	Violation of algorithms or methods by users intentionally or unintentionally causes failure by societal wellbeing.	Failure to societal well-being by users violation
Users and system interphase	Environmental Wellbeing	User protocols or definitions missuses	Violation of algorithms or methods by users intentionally or unintentionally causes failure by environmental wellbeing.	Failure to environmental well-being by users violation
Users and system interphase	Human Agency and Oversight	User protocols or definitions missuses	Violation of algorithms or methods by users intentionally or unintentionally causes failure by Human agency and oversight.	Failure to human agency and oversight by users violation
Users and system interphase	Privacy	User protocols or definitions missuses	Violation of algorithms or methods by users intentionally or unintentionally causes failure by privacy.	Failure to privacy by users violation
Users and system interphase	Bias	User protocols or definitions missuses	Violation of algorithms or methods by users intentionally or unintentionally causes failure by bias.	Failure to bias by users violation
Users and system interphase	Users Values	User protocols or definitions missuses	Violation of algorithms or methods by users intentionally or unintentionally causes failure by users' values.	Failure to users values by users violation
Users and System Interphase	Privacy and Data Governance	Individuals information disclosure violation	Inappropriate disclosure of personal data internally within your organisation due to a lack of appropriate controls being in place	Failure to privacy by lack of internal control
Physical driver	Privacy and Data Governance	Individuals’ information disclosure violation	Accidental loss of electronic equipment by personnel	Failure to the disclosure of personal information