From Blade Runners to Tin Kickers: what the governance of artificial intelligence safety needs to learn from air crash investigators

Prof Carl Macrae
Nottingham University Business School
University of Nottingham
carl.macrae@nottingham.ac.uk
@CarlMacrae

6.10.21

What should we do when artificial intelligence (AI) goes wrong? AI has huge potential to improve the safety of societally critical systems such as healthcare and transport, but it also has the potential to introduce new risks and amplify existing ones. For instance, biases in widely deployed diagnostic AI systems could adversely affect the care of a large number of patients (Fraser et al. 2018), and hidden weaknesses in the perception systems of autonomous vehicles may regularly expose road users to significant risk (NTSB 2019). What are the most appropriate strategies for governing the safety of AI-based systems? One answer emerges from taking contrasting looks forwards to our imagined dystopian AI future and backwards to the progressive evolution of aviation safety.

In science fiction, one of the most iconic portrayals of the control of errant AI is presented in Ridley Scott's (and more recently Denis Villeneuve's) *Blade Runner*. Based on Philip K Dick's dystopian novel (1968), professional 'Blade Runners' track down rogue humanoid AI systems which they violently 'retire' (destroy, or perhaps kill). Images of hazardous AI that adaptively escapes the tight confines of human control and must be covertly pursued and punitively dismantled are a common motif in popular culture (Cave and Dihal 2019). And, while fictional, the work of a Blade Runner offers an extreme illustration of one approach to

the governance of AI safety. The principles underlying what might be termed 'Blade Runner governance' of AI safety have four key characteristics. First, it is *atomised* and focuses on identifying, disabling and removing a deviant individual or subsystem. Second, it is *punitive* and employs correctional strategies that seek accountability and retribution for prior behaviour. Third, it is *compliance-oriented* and focuses on deviant behaviour that breaches some pre-defined standard. Fourth, it is *closed* and operates through intentionally covert, hidden or secret processes.

In this imagined future Blade Runners are tasked with pursuing particularly sophisticated rogue AI, but these underlying governance principles are far from fictional—they are already apparent in response to the failures and risks of current AI systems, with individual human operators blamed for the failure of complex AI (Levin 2016; 2020; Elish 2019), and a profusion of AI ethical guidelines that frame accountability as a retrospective process of determining responsibility for past failure (Jobin et al. 2019).

The problem is that these principles are contrary to much of what we know about how to improve safety in complex sociotechnical systems. A more productive and practical image to guide the governance of AI safety is not that of the Blade Runner, but is rather more prosaic, less familiar but much better understood—that of the 'Tin Kicker': air crash investigators who 'kick tin' on accident sites while picking over wreckage (Byrne 2002; Nixon and Braithwaite 2018). Professional accident and safety investigators have been central to the continuous improvement of flight safety since the dawn of aviation (Macrae 2014). The first independent air crash investigation was conducted in 1912 (Hradecky 2012), followed soon after by the establishment of the UK's accident investigation body in 1915 (AAIB 2021). Professional safety investigation agencies have since become common in many transport sectors around the world and are emerging in other safety-critical industries like healthcare (Macrae and Vincent 2014; 2017).

AI accident investigation will be critical for building trust in AI and ensuring that AI failures are widely learnt from (Winfield and Jirotka, 2017; 2018; Winfield et al. 2021)—and these investigative activities will need to grapple with risks arising from the inherently sociotechnical nature of AI systems (Macrae 2021). But more fundamentally, the principles and practices that have guided the work of 'Tin Kickers' for over a century offer important foundational lessons for the governance of AI safety. What might be termed 'Tin Kicker governance' of AI safety illustrates a dramatic counterpoint to the principles of the AI Blade Runner. Rather than focusing on individual elements, it emphasises the *systemic* nature of risk, drawing on analytical methods and models that capture the complex sociotechnical processes that shape deviations in expected behaviour at all system levels, from technological to regulatory (ATSB 2007; Waterson et al. 2017). Rather than seeking accountability for past failures, it is exclusively *learning-oriented* and purposefully does not attribute liability or blame but instead seeks to create active accountability for future improvement (Braithwaite 2011). Rather than addressing compliance with accepted standards, it focuses on understanding the *practical realities* of complex systems, and why unexpected or deviant behaviours may be situationally rational and adaptive given particular contexts, constraints and affordances (Macrae 2014). And rather than a closed and covert process, it is fundamentally *participatory*, openly engaging with all relevant stakeholders to collaboratively understand reasons for failure and develop appropriate recommendations—whilst retaining authority over those findings and recommendations (Macrae and Vincent 2014).

The work of Tin Kickers, and the principles that guide this work, therefore offer a rich and productive exemplar that holds important lessons for the development of more effective strategies of AI safety governance—and offers a stark contrast to the model of an AI Blade Runner that exists in the popular imagination. Indeed, Tin Kickers are already at work in AI safety, investigating the failures of some self-driving cars and automated driving systems (NTSB 2019; 2017). This work has begun to reveal some of the organisational inertia and

cultural blindspots that will need to be addressed to establish more systemic, learning-oriented and participatory approaches to AI safety governance. Tesla failed to even acknowledge federal investigators' recommendations following a fatal 2016 crash (O'Kane 2020), while Uber responded to its fatal self-driving crash by committing to build a safety management system for its test vehicles within 5 years (NTSB 2019)—a timeline seemingly as long as its manufacturing partner's projection for large-scale deployment of self-driving cars (O'Kane 2019; Volvo 2019). These early challenges serve to emphasise the urgent importance—rather than the impossibility—of creating systems of AI safety governance that embody the principles of Tin Kickers, long before we need to resort to those of the Blade Runner.

References

AAIB (2021) About us. Air Accidents Investigation Branch. https://www.gov.uk/government/organisations/air-accidents-investigation-branch/about

ATSB (2007) Analysis, Causality and Proof in Safety Investigations. Canberra, Australian Transport Safety Bureau

Braithwaite J (2011) The Essence of Responsive Regulation. UBC Law Review 44(3):475-520

Byrne G (2002) Flight 427: Anatony of an Air Disaster. London, Springer

Cave S, Dihal K (2019) Hopes and fears for intelligent machines in fiction and reality. Nat Mach Intell 1:74–78

Dick PK (1968) Do Androids Dream of Electric Sheep? London, Doubleday

Elish MC (2019) Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction. Engaging Science, Technology, and Society 5:40–60

Fraser H, Coiera E, Wong D (2018) Safety of patient-facing digital symptom checkers. Lancet 392(10161):2263–2264

Hradecky S (2012) United Kingdom's Air Accident Investigation Board celebrates 100 years of air accident investigation. The Aviation Herald. http://avherald.com/h?article=450d2364&opt=1

Jobin A, Ienca M, Vayena E (2019) The global landscape of AI ethics guidelines. Nat Mach Intell 1:389–399

Levin S (2016) Uber blames humans for self-driving car traffic offenses as California orders halt. The Guardian, 15th Dec 2016.

https://www.theguardian.com/technology/2016/dec/14/uber-self-driving-cars-run-red-lights-san-francisco

Levin S (2020) Safety driver charged in 2018 incident where self-driving Uber car killed a woman. The Guardian, 16th Sep 2020. https://www.theguardian.com/us-news/2020/sep/16/uber-self-driving-car-death-safety-driver-charged

Macrae C (2014) Close Calls: Managing Risk and Resilience in Airline Flight Safety. London, Palgrave Macmillan

Macrae C (2021) Learning from the Failure of Autonomous and Intelligent Systems: Accidents, Safety and Sociotechnical Sources of Risk.
SSRN. https://ssrn.com/abstract=3832621

Macrae C, Vincent C (2014) Learning from failure: the need for independent safety investigation in healthcare. J Roy Soc Med 107:439–43

Macrae C, Vincent C (2017) A New National Safety Investigator for Healthcare: The Road Ahead. J Roy Soc Med 110(3):90-92

Nixon J, Braithwaite GR (2018) What do aircraft accident investigators do and what makes them good at it? Developing a competency framework for investigators using grounded theory. Saf Sci 103:153–161

NTSB (2017) Collision between a car operating with automated vehicle control systems and a tractor-semitrailer truck near Willliston, Florida, May 7 2016. Washington DC, National Transportation Safety Board

NTSB (2019) Collision between vehicle controlled by developmental Automated Driving System and Pedesterian, Tempe, Arizona, March 18, 2018: Accident Report NTSB/HAR-19/03. Washington DC, National Transportation Safety Board

O'Kane S (2019) Uber debuts a new self-driving car with more fail-safes. The Verge. https://www.theverge.com/2019/6/12/18662626/uber-volvo-self-driving-car-safety-autonomous-factory-level

O'Kane S (2020) Tesla ignored safety board's Autopilot recommendations, chairman says. The Verge. https://www.theverge.com/2020/2/25/21152984/tesla-autopilot-safety-recommendations-ignored-ntsb-crash-hearing

Volvo (2019) Volvo Cars and Uber present production vehicle ready for self-driving. Volvo Cars Global Newsroom. https://www.media.volvocars.com/global/en-gb/media/pressreleases/254697/volvo-cars-and-uber-present-production-vehicle-ready-for-self-driving

Waterson P, Jenkins DP, Salmon PM, Underwood P (2017) 'Remixing Rasmussen': The evolution of Accimaps within systemic accident analysis. Appl Ergon 59(Part B):483–503

Winfield AFT, Jirotka M (2017) The Case for an Ethical Black Box. In: Gao Y, Fallah S, Jin Y, Lekakou C. (eds) Towards Autonomous Robotic Systems. TAROS 2017. Lecture Notes in Computer Science, vol 10454. Springer, Cham. https://doi.org/10.1007/978-3-319-64107-2_21

Winfield AFT, Jirotka M (2018) Ethical governance is essential to building trust in robotics and artificial intelligence systems. Philos T Roy Soc A 376(2133):20180085

Winfield AFT, Winkle K, Webb H, Lyngs U, Jirotka M, Macrae C (2021) Robot Accident Investigation: A Case Study in Responsible Robotics. In: Cavalcanti A, Dongol B, Hierons R, Timmis J, Woodcock J (eds) Software Engineering for Robotics. London, Springer.