Three Principles for Testing Safety-Critical Robotics and AI In Public

Jason Millar[1]
Caitlin Heppner[2]

INTRODUCTION

We have recently emerged from truly desperate times. With the global COVID-19 pandemic largely in our rear-view mirror, it is surprisingly difficult to look back and clearly recall how, for what felt like an eternity, our freedoms were shelved while we waited for a vaccine that would solve all our problems. There is no need to replay the details of that collective desire for a technological fix—you have likely done that with loved ones during countless post-pandemic get-togethers. But it is worth recalling, for the purpose of the argument we make in this paper, that even in those most desperate times we clung to a set of principled practices designed to help us find safe vaccines that we knew with some certainty posed few enough risks to individuals that we considered them acceptable to release for widespread use.

That we would have considered it grossly irresponsible to perform safety testing on candidate vaccines (or other candidate pharmaceuticals) after, or in parallel with, releasing them for widespread use, is a testament to the hard-won gains we have made in medical ethics in the past fifty odd years. Our understanding of the *safety-critical* nature of molecular technologies—the fact that using them poses significant risks to people's physical and/or mental health—is at the core of that principle. Coupled with histories of irreparable, often shameful, harms delivered to masses of trusting and unsuspecting people at scale, we have developed rigorous ethical testing and reporting protocols to gate the release and subsequent widespread use of those technologies. We are right to have done so, and ought to keep a watchful vigilant eye on the gates that protect us.

In this paper we argue that some robotics and artificial intelligence (RAI) applications are safety-critical in the same sense that some molecular technologies, such as vaccines, are safety-critical. Put plainly, we know that some RAI applications pose inherent risks to people's physical and/or mental wellbeing. Yet, in our current technological milieu, safety-critical RAI technologies are regularly and commonly tested at scale, in public, on trusting and unsuspecting people, with few or no protections in place to ensure public safety. We have good reason to question these commonplace testing practices and ask how we can do better.

Principles we have developed in bioethics to govern safety testing for safety-critical molecular technologies and other experimental medical interventions have served us well in their respective domains. Thus, we argue further that we can, and ought to, borrow and apply some testing and reporting

[1] Jason Millar is the Canada Research Chair in the Ethical Engineering of Robotics and Artificial Intelligence, and Associate Professor, at the School of Engineering Design and Teaching Innovation, Faculty of Engineering, University of Ottawa. He holds a cross-appointment in the Philosophy Department. jmillar@uottawa.ca
[2] Caitlin Hepner is a PhD student in the Philosophy Department at the University of Ottawa.

principles from bioethics to RAI, to make RAI testing safer and more trustworthy for the public. Namely, we argue that testing safety-critical RAI in public should, at a minimum: 1) involve more robust informed consent that alerts unsuspecting participants to the fact that they are the subjects of technological experiments that undeniably expose them to harm; 2) adopt the equivalent of a "provision of standard of care" requirement to ensure existing technological safeguards remain in place to protect human subjects; and, 3) include more robust adverse event reporting to support public scrutiny, informed consent practices, and experiment termination practices.

Our argument is not a call for randomized control trials (RCTs) for RAI, nor is it a call for imposing onerous pre-release testing on all RAI developers. It is also not a call to treat RAI as equivalent to molecular technologies in all applications. Our argument is a call to recognize the similarities between two sets of technologies and ask why it is that we adhere to principles (even in our most desperate times) to protect test subjects from harm in one case, and not in the other. It proposes three tried and tested principles that could help protect those people who are currently at risk of harm.

DEFINING SAFETY-CRITICAL TECHNOLOGY

Why introduce the notion of a "safety-critical" technology? Surely, one might fairly assume, engineers and computer scientists are adequately focused on safety when they design RAI applications. Indeed, they are very focused on safety. We do not wish to suggest that technologists wantonly disregard people's safety. Yet, despite their regular focus on safety, the current operationalized technological conceptions of "safety"—the scope of the term, and the practices it engenders, which developed within professional and sociotechnical contexts—do not consistently anticipate, acknowledge, or adequately prevent the full range of harms posed by current RAI innovation practices (Salem et al, 2023). Those concepts and practices, for example, allow testing risky RAI on trusting and unsuspecting people, causing them harm. We thus need a new technological category that captures technologies that are not specifically designed as safety systems/technologies, and that recognizes the significant risks inherent to some RAI technologies.

According to their manufacturers' own classifications, the RAI applications that provide the case studies in this paper—automated driving systems and social media algorithms (SMAs)—are not all technically classified as safety systems. Auto manufacturers tend to classify their Advanced Driver Assist Systems (ADAS), such as lane keep assist and "hands-free driving" systems, as convenience features (Cadillac, 2023; Tesla, 2023b), while SMA developers like Facebook describe their platforms as neutral egalitarian content hosts (Gillespie, 2010).[3] These characterizations obfuscate the risks posed by the underlying technologies. Thus, rather than relying on corporate fiat, we need to shift conceptions of safety and the practices they encompass to ones that are anchored in RAI's objective safety impacts.

There is currently no single definition of what constitutes a "safety-critical" technology. Knight (2002) suggests that "safety-critical systems" can been classified in engineering as "systems whose failure might endanger human life, lead to substantial economic loss, or extensive environmental damage" (p. 547). However, the scope of this definition, if taken as an indication of systems about which we ought to be particularly concerned when designing and deploying them, or governing their testing, limits our attention in two problematic ways. First, it distracts us from the effects a technology might have on people using it

[3] Gillespie's analysis of the term "platform" is useful here. "calling their service a 'platform' can be a way not to trumpet their role, but to downplay it. . . . In the effort to limit their liability not only to those legal charges by also more broadly to the cultural charges of being puerile, frivolous, debased, etc., intermediaries like YouTube need to position themselves as just hosting – empowering all by choosing none" (p. 357).

as intended—only failure-related harms are considered. Second, it narrows the category of relevant failure-related harms to those that that are physical, economic, or environmental in nature and substantial in effect. A broader definition is required if we are to account for a fuller range of harms, including those less severe yet worthy of our attention, and those resulting from either intended use or failure.

According to Salem et al. (2022), from "a more general [technical standards] perspective, the term 'safety' tends to refer to a state in which there is an absence of unreasonable risk" (para. 10), despite there being no commonly used definition of the term. This gets us closer to a workable definition as it hinges on any unreasonable risk, thus broadening the definition enough to capture any harms that, upon reflection, we should be concerned with. But "safety" here describes a state of affairs rather than a category of things.

Hence, a safety-critical technology, as we suggest in our introduction, is simply a technology that, when used, poses significant risks to people's physical and/or mental wellbeing.[4] A safety-critical technology is thus very different from a safety technology or system; safety systems are designed with safety in mind, whereas safety-critical technologies pose inherent risks to people's wellbeing regardless of their designers' intention. Borrowing further from Salem et al. (2023), we would want to develop governance mechanisms for those safety-critical technologies that expose people to *unreasonable* risks to their wellbeing.

SAFETY-CRITICAL RAI IS BEING TESTED IN PUBLIC

Uber's Automated Vehicle Pedestrian Fatality

In 2018 Uber was testing a fleet of fully autonomous test vehicles on Arizona's public roads. One of those vehicles struck and killed Elaine Herzberg, a pedestrian who was crossing the road. The vehicle involved was a Volvo XC90, which comes equipped with a standard collision detection and emergency braking system. Uber's autonomous driving software "saw" Herzberg crossing the street but failed to successfully classify her as a pedestrian and ultimately failed to brake. The safety driver in the vehicle, who was distracted, also did not notice Herzberg and failed to brake in time. Following Herzberg's death, the safety driver was charged with negligent homicide, ultimately pleading guilty to endangerment (Shepardson, 2023). Uber was not charged.

There were many contributing factors to this accident, which marked the first time a pedestrian was killed by a vehicle driving in autonomous mode, but one is of particular relevance to our argument. The resulting investigation revealed that Uber had disabled Volvo's emergency collision detection and braking system—among other standard factory safety systems—to test their own experimental autonomous driving software on public roads (Atiyeh, 2019). Uber's explanation for their decision was that they disabled the featured to "reduce the potential for erratic vehicle behavior" (Wakabayashi, 2018, para. 6). According to then chief research officer at the Insurance Institute for Highway Safety, "I think it's possible that, had the system been able to intervene, the fatality may not have occurred" (Ganz, 2018, para. 3).

Tesla's Perpetual Automated Driving Software Beta Testing

A quick glance through the Autopilot and Full Self-Driving sections of the Tesla Model 3 Owner's Manual makes it clear that many, if not all, of those automated driving systems are (and have been for years) "beta" versions (Tesla, 2023a). The software industry uses the "beta" designation to indicate that a piece of software is almost ready for full commercial release, but that the software needs to be tested with a large group of users under real use conditions (PCMag, 2023). There is an explicit understanding that beta software is still buggy, and could cause problems, including the loss of data. Since they have been in

[4] We narrow our definition here to exclude unreasonable economic and environmental harms only to keep this paper manageable. They should also be the subject of governance measures intended to prevent or mitigate them.

perpetual "beta" mode since their introduction, most of Tesla's automated driving features are, by definition, being *tested* on public roads.

It is unclear the extent to which Tesla drivers fully appreciate the still-experimental nature of beta version software, or if they understand the role they are playing in those public tests every time they engage those systems, or the risks that poses to themselves, other drivers, pedestrians, and other vulnerable road users. To be clear, and in contrast to the Uber case described above, evidence suggests those drivers have received no formal training how to operate those systems appropriately (Szilagyi et al., 2023). According to recently released data, Tesla's automated driving systems have been implicated in "736 crashes since 2019, including 17 fatalities" (Blanco, 2023, para. 2). The National Highway Traffic Safety Administration (NHTSA) has automated driving incident data indicating that of the 807 reported automation-related crashes, almost all involved Teslas—Subaru coming in second with 23 (Blanco, 2023).

Facebook's Emotional Contagion Study

In 2014, researchers at Cornell University published a study titled "Experimental evidence of massive-scale emotional contagion through social networks," which used data from one of Facebook's massive (*N*=689,003) emotional manipulation experiments. The experiment manipulated the degree to which users were exposed to positive and negative expressions on their news feed by reducing the number of positive posts on user's news feeds in one group, and the number of negative posts on users' news feeds in a second group. Based on an analysis of the affective content in users' posts after the parallel experiments were conducted, Kramer et al. (2014) propose that positive and negative emotional states can be "transferred" based on text-based computer-mediated communication. In other words, they suggest that emotional states can be *induced* by social media feeds and their algorithms.

When criticism was levied against the researchers and Facebook for their data collection practices in this study, the authors claimed that Facebook's Data Use Policy, which users were required to agree to prior to creating their accounts, constituted informed consent for emotional manipulation experiments. This claim has been challenged by Flick (2016) who argued that the study neglected to obtain informed consent from participants, proposing that Responsible Research and Innovation frameworks (RRI) could provide the groundwork for bridging the ethical oversight gap between university and industry research.

Instagram's Sticky Harms

In 2021 whistleblower Francis Haugen released internal documents to *The Wall Street Journal* demonstrating that Meta's internal research had established a correlation between Instagram use and negative mental health impacts among teenage users (Wells et al., 2021). Meta had been studying the impact of their product since 2019 and, according to their own internal documents, which had been kept secret, "thirty-two per cent of teen girls said that when they felt bad about their bodies, Instagram made them feel worse" (Wells et al., 2021, para. 5). According to Gayle (2021),

Among the most concerning findings was that among users who reported suicidal thoughts, 13% in the UK and 6% in the US traced them back to Instagram. Another transatlantic study found more than 40% of Instagram users who reported feeling "unattractive" said the feeling began on the app; about a quarter of the teenagers who reported feeling "not good enough" said it started on Instagram. (para. 9)

Evidence suggests that social media use leads to an increase in mental illness and mental harms in young users—an increase in body dysmorphia, body dissatisfaction, anxiety, depression, and eating disorders. Although spanning gender identities, these harms did arise most significantly in female-identifying youth

(Kelly et al., 2018).[5] In short, social media platforms, and the AI that powers them at scale, are known to cause harms that have proved to be at times extreme, long-lasting, and potentially fatal for individual users (Holland & Tiggemann, 2016).

The AI at the heart of social media platforms, which decides which information to serve to each user, has long been known to cause other harms. Algorithmic bias has been studied in depth by Noble (2018) and Benjamin (2019), who have shown that algorithmic bias leads to an increase in racialized and gender-based harms. Filter bubbles and echo chambers have featured predominantly in academic discourse regarding online political and democratic harms, with varying degrees of consensus over both phenomena (Bozdag & van den Hoven, 2015).[6] However, the impact of streamlined algorithmic feeds, filter bubbles, and echo chambers on more vulnerable youth populations cannot be ignored.[7] Furthermore, these platforms are intentionally designed to be addictive, deploying sophisticated psychological manipulation tactics (such as the emotional contagion study described above) to achieve that effect (Schwär, 2021), thus compounding the harms by making them "sticky". Social media platforms' addictive quality has been compared to that of addictive painkillers as it operates on similar neural mechanisms (Schwär, 2021).

What Do We Owe the Subjects of Safety-Critical RAI Testing?

As described above, testing safety-critical RAI in public puts not only the direct users at risk, but also poses risks to a large body of indirect users. In addition to vehicle occupants, automated vehicle testing on public roads puts other drivers, pedestrians, cyclists, and other vulnerable road users in harm's way. We are also beginning to understand how SMAs impact not only those feeding and consuming the SMA ecosystem, but additionally impact the political and social landscape, resulting in population-level risks. Social media platforms can no longer be treated as neutral platforms, because we know they create a novel information ecosystem that has been shown to impact elections, erode democracy, and provide the framework for the promulgation of hate groups (Benjamin, 2019; Frenkel et al., 2018; O'Neil, 2016; Noble, 2018). As more RAI applications are deployed and tested in public at scale, we are pressed to recognize how allowing those practices is simply inconsistent with our ethical posture toward testing other safety-critical technologies, such as pharmaceuticals. The upshot is that safety-critical RAI test subjects—most of whom do not understand their role as such—who we find at elevated risk due to their being subjected to public testing, deserve protections.

REGULATING HARMS

Generally, in both Canada and the United States, harmful substances and products are regulated by law, predicated on reducing harm to the public. A brief glance at the regulation of harmful products in both Canada and the United States, including tobacco, alcohol, cannabis, narcotics, and medications, provides

[5] This analysis of the UK Millennium Cohort study associated similar health issues between boys and girls, with female-identifying respondents reporting larger symptoms. This study also linked extended periods of use with an increase in mental health symptoms.

[6] Bozdag & van den Hoven (2015) suggest that filter bubbles limit freedom of choice, minimize the crucial pool of information required for deliberative democracy, limit access to diverse thought, and block reliable channels of information; alternatively, see Axel Burns (2019) challenges the usefulness and accuracy of this work, arguing that the concept of the filter bubble and its captivation has only redirected scholarly attention away from more pressing topics.

[7] Consider the study conducted by Fardouly et al. (2018) in which increased social media use and more specifically an increase in the frequency of viewing fitspiration images on social media is suggested to be linked with an increase in body dissatisfaction and self-objectification. Fardouly et al. (2018) supports the study by Tiggemann and Zaccardo (2015), which found that fitspiration images on Instagram increased body dissatisfaction and negative moods among young women.

a partial, but useful, conceptual foundation for understanding where some of the lines are currently drawn between technologies that are heavily regulated, and those that are not. It also helps to understand what considerations are relevant when considering whether additional classification and regulation is needed for RAI systems posing similar harms.

Tobacco, considered by the World Health Organization [WHO] to be the "leading cause of death, illness and impoverishment" (WHO, 2021a) and linked to the deaths of half its users, with more than 8 million deaths attributed to the substance, is regulated by the Tobacco and Vaping Products Act in Canada, which regulates the manufacturing, sale, labelling and promotion of all tobacco products. This legislation limits nicotine concentration (20mg/mL in vaping products), bans flavouring and sweeteners, energy and vitality substances, and colouring (*Tobacco and Vaping Products Act*, SC 1997, c 13). Access is also regulated: tobacco products can only be sold to individuals over the age of 19 (*Tobacco and Vaping Products Act*, SC 1997, c 13). In the United States, tobacco products are regulated by the Food and Drug Administration (FDA) which restricts the manufacturing, distribution, and marketing of all tobacco products (Family Smoking Prevention and Tobacco Control Act, 2009). Alcohol is known to cause similar harms since alcohol consumption has been associated with 3 million deaths yearly, a range of mental and behavioural disorders and injuries, incidence of infection diseases, and social and economic loses at both the individual and the societal level (WHO, 2018). While alcohol legislation in Canada varies from province to province (like tobacco regulations), alcohol regulations in Canada include labelling and packaging requirements, laboratory testing, pricing baselines, regulated sales and distribution, marketing, promotion and sales requirements and limitations, and licensing requirements. Regulation of alcohol in the United States is largely state dependent; individual states largely set the manufacturing standards, the regulations for the sale and promotion of alcohol, and are responsible for managing alcohol-related problems.

Similar harms arise in the examination of cannabis use: memory, attention, and coordination are impacted, and may be impacted for up to 24 hours after consumption; acute health effects include impairment of cognitive development and impairment of psychomotor performance; chronic health effects include impairment of cognitive function, cannabis dependence, exacerbated symptoms of schizophrenia, and physical damage to the mouth, throat, lungs, and pulmonary system, including an increased risk of chronic bronchitis (WHO, 2016). Despite legalization of cannabis in Canada in 2018, there remains strict laws in Canada which regulate the sale, distribution, access, promotion, packaging, labelling, and display of cannabis. The minimum age for cannabis use is 18, with restrictions on the amount of cannabis an individual person is allowed to possess in public. The packaging and promotion of cannabis is also regulated similarly to tobacco. In the United States, the control and legality of cannabis varies, with regulation ranging from complete legalization in states such as California and Vermont to fully illegal in states such as South Carolina, Kansas, Wyoming, and Idaho.

Lastly, opioids and related pharmaceuticals have been implicated in 500,000 deaths worldwide, with approximately 36.3 million people identified with a substance use disorder in 2019 (WHO 2021b). The risk of overdose is high, and while recovery from addiction is possible, the WHO estimates that only about 10% of those with a substance use disorder receive treatment (WHO, 2021b). For all pharmaceutical production, international norms and standards for the production of pharmaceuticals are developed through an expert collaboration between the WHO Expert Advisory Panel on the International Pharmacopoeia and Pharmaceutical Preparations and various international and national health authorities including regulatory authorities, health agencies, industry specialists, national institutions, and nongovernmental organizations (WHO, n.d.).

With this brief overview of harmful product regulation in both Canada and the United States, we observe that only some harmful products and some kinds of harms are regulated. Perhaps the boundaries of the regulation of harms are more concretely illustrated by considering the lack of regulation for so-called "low-risk products," which are subject to little or no regulation because they are *not* shown to be harmful. Under current US FDA regulations, health related products—intended to diagnose, cure, mitigate, prevent, or treat disease or conditions—are regulated by Federal Food, Drug, and Cosmetic Act, while general wellness products that are simply intended to promote a healthy lifestyle are considered low-risk, and are therefore not subject to the same rules (Federal Food, Drug, and Cosmetic Act, 2022; FDA, 2019). For example, manufacturers of low-risk products are not required to adhere to registration, listing, and premarket notification requirements, labelling requirements, manufacturing requirements, or Medical Device Reporting requirements (FDA, 2019). The FDA's Centre for Devices and Radiological Health (CDRH) defines these products by two qualities: products that "(1) are intended for *only* general wellness use, as defined by their guidance, and (2) products which present a low risk to the safety of users and other persons" (FDA, 2019, p. 2). If the product is shown to be invasive, intended to be implanted, or "involves an intervention or technology that may pose a risk to the safety of users and other persons if specific regulatory controls are not applied" (FDA, 2019, p. 5), the product does not meet the threshold to qualify as low-risk.

According to the current FDA recommendation, the decision whether to consider an RAI application low- or high-risk, and thus, the extent to which it ought to be regulated, hinges on a determination of the nature of its "risk to the safety of users and other persons" (FDA, 2019, p. 2).

What can be said of the risks associated with safety-critical RAI as described above? None of those harms would be considered mere inconveniences; the various RAI harms described above, if considered side-effects of using the technology, are severe.

PRINCIPLES FOR TESTING SAFETY-CRITICAL AI SYSTEMS

Even if we choose to classify safety-critical RAI as low-risk, thus deeming them outside the scope of existing regulatory scrutiny, it is undeniable that they can pose significant or elevated risks when tested at scale in public. It is therefore reasonable to develop and apply a framework consisting of principles to mitigate those risks. At this time, there are no regulatory bodies for SMA testing and the regulations governing automated driving systems are quite limited. Within the medical and pharmaceutical industries, governing bodies include university ethics boards, publishing ethics boards, funding ethics boards, professional practice bodies, and government oversight committees. There are regulations which guide testing practices, including informed consent, post-trial responsibilities, the provision of standard of care, and adverse event reporting. Additionally, Health Canada provides some regulation of medical devices sold in Canada through its mandate to ensure product safety and efficacy (Da Silva et al., 2022). As we described above, the regulation of harmful substances (i.e. molecular technologies) stems from the risks and associated harms those technologies pose to the public. We have described how RAI—automated driving systems and social media algorithms—pose similar risks and associated harms. Yet, we consider it unthinkable to test molecular technologies in public, and so impose strict restrictions on how that testing ought to be appropriately conducted, while we allow relatively unfettered testing of RAI on trusting and unsuspecting publics.

To address this asymmetry and the mitigate potential harms, we propose a basic ethical framework for testing safety-critical RAI systems in public. Our proposed framework borrows three well-established

principles from bioethics and pharmaceutical testing which, if adapted appropriately to RAI testing practices, can improve public safety. The three principles are informed consent, the provision of standard of care, and adverse event reporting.

Informed Consent

The Nuremburg Code (1947) was established in reaction to the abusive medical procedures performed by Nazi physicians on prisoners in concentration camps in World War II. As an international statement of principles meant to govern all medical research on human beings, it stipulates that "the voluntary consent of the human subject is absolutely essential" (Nuremburg Code, 1947, para. 2). This means that the person involved should have the legal "capacity to give consent; should be situated as to be able to exercise the free power of choice… and should have sufficient knowledge and comprehension of the elements of the subject matter involved so as to enable him to make an understanding and enlightened decision" (Nuremburg Code, 1947, para. 1). The necessity of informed consent is predicated upon the general right to individual self-determination and is instrumentally valuable for protesting patient autonomy. For an action to be autonomous under Beauchamp and Childress's (2008) conception, which is now commonplace in medical practice, the decision only needs to be made with a substantial degree of understating and voluntariness.

The requirement of informed consent has long been problematized within bioethics: at what times can an individual be said to be capable of proper consent,[8] to what degree can an individual be informed,[9] where do physician biases impact the delivery of important information,[10] and many other situations which arise that complicate a researcher's ability to fully receive "informed consent" from test subjects. Nonetheless, the principle of informed consent as it currently is used to guide and shape research (i.e. testing on human subjects) provides concrete and clear terms by which researchers *must* inform their test subjects and receive consent. In Canada, the guidelines which govern all federally funded research state that research on human subjects cannot be carried out unless free and informed consent is obtained from the competent subject before and throughout the research project.

Informed consent, applied to safety-critical RAI testing (e.g. Meta's emotional contagion experiment) would require, at a minimum: informing subjects that they are, indeed, subjects of a technological experiment; describing to potential subjects the nature and purpose of the test; explaining to potential subjects that there are foreseeable risks associated with participation; and offering them a meaningful opportunity to opt-out if they so choose.[11]

The Provision of Standard of Care

It is standard practice in medical research, in line with the Declaration of Helsinki (1964; 2013), to test new interventions against best proven interventions, with few exceptions (World Medical Association, 2013). This requires that test subjects *always* be provided the most effective treatment available. A patient cannot be denied the existing standard of medical care in favour of testing a novel intervention, unless

[8] Consider the cases of children desiring medical care without parental consent or the state-of-mind of a late-stage Alzheimer's patient.

[9] Challenges arise here where most patients do not have the education to fully comprehend diagnosis or treatment options.

[10] Consider the difference between the information a physician in support of abortion or MAID might provide a patient vs. the information a physician opposed to abortion or MAID might provide. Some have demonstrated that it is unfair or even impossible to have a physician provide unbiased/politically neutral information to their patients.

[11] We note that the language typically contained in today's end user licensing agreements would not satisfy informed consent requirements.

"available evidence suggests that the intervention will be at least as advantageous, in the light of foreseeable risks and benefits, as any established effective alternative" (Council for International Organizations of Medical Sciences, 2016, p. 9). Studies must be conducted in such a way as to benefit the scientific community without "delaying or withholding established effective interventions from participants" (Council for International Organizations of Medical Sciences, 2016, p. 15). This regularly poses challenges to the advancement of medical knowledge, and at times requires trials to be put on hold, even indefinitely. Nonetheless, the research community upholds that, except under exceptional circumstances, test subjects must receive the most effective standard care available, regardless of the potential benefit of the experimental intervention.

We can see how this principle applies to testing safety-critical RAI in public using Uber's pedestrian fatality as an example. Volvo's collision detection and emergency braking system was the "standard of care" in that driving context. Uber disabled it, thus subjecting its human participants—the other drivers, pedestrians, cyclists, other vulnerable road users, and ultimately Elaine Herzberg—to an experimental system. Elaine Herzberg was the victim of a technological experiment gone bad. Regardless of the difficulties Uber or Tesla might experience bringing their automated vehicles to market, those test subjects should be afforded the benefits of existing, validated, vehicle safety features.

Robust Adverse Event Reporting

Market Authorization Holders (MAHs) are required to report serious adverse reactions (expected and unexpected) which involve their marketed health products in accordance with the Food and Drugs Act and its regulations. Adverse reactions are characterized by the fact that a causal relationship between the drug and the occurrence is *suspected*. In the regulations, serious adverse reactions are defined as noxious, unintended responses to a drug that occurs at any dose and that result in "in-patient hospitalization or prolongation of existing hospitalization, causes congenital malformation, results in persistent or significant disability or incapacity, is life-threatening or results in death" (*Food and Drug Regulations*, 2023, section C.05.001). In practice, all manner of side effects are reported to possible consumers, including some very minor, possibly unnoticeable side effects like dry mouth or itchy skin.

Meanwhile, despite evidence and knowledge of risks, RAI producers are not required to warn users of them. These RAI "side-effects" include those described in context of automated vehicles above, as well as those we have discussed that are associated with the use of Instagram and other social media platforms. There are more. Drivers using automated or driver assist systems might find their inattention (e.g. eyes-off-road) increasing by 33%, and they might stop paying attention to the road for upwards of 45 seconds at a time (Llaneras et al., 2013; Gaspar & Carney, 2019).

How might the standard of adverse event reporting be included in the testing and release of new safety-critical RAI applications? One possibility is to use adverse event reporting as the basis of industry standards to halt testing if harms exceed minimal agreed-to thresholds. Another would be to provide publicly available data on the side effects of using a particular RAI application to underwrite robust informed consent practices.

CONCLUSIONS

We began this paper reflecting on the strength of our collective convictions for applying principles when testing some safety-critical technologies for public consumption. In the depths of the COVID-19 pandemic, we held to principles that slowed the development and rollout of vaccines to ensure their safety. Those principles are predicated upon a general understanding that the use of certain technologies poses significant risks to people's physical and/or mental health. We do not apply those principles evenly across

technological domains—as we have described, safety-critical robotics and AI technologies are regularly tested in public, exposing trusting and unsuspecting people to significant risks.

Established ethical testing practices have historically helped mitigate harms, so we have proposed applying three of the guiding principles common to those practices—informed consent, the provision of standard of care, and robust adverse event reporting—to inform the testing of safety-critical RAI. We have not specified the means by which to apply, or enforce, those principles; they could be applied as industry best-practices, standards, or required by law. We recommend them here because they are purpose-built to protect against the very harms that have and continue to threaten people exposed to the public testing of RAI, which we know to be risky.

Our argument is not intended to develop overly cumbersome testing requirements, but instead offer tried and tested means of ensuring public safety in cases where the practical development of technologies requires testing in public. By adopting the core principles of informed consent, the provision of standard of care, and adverse event reporting, and demonstrating their possible applications in the testing of RAI, we have offered three possible means by which industry can meet the necessary testing requirements for expedient development while meeting safety and ethical standards that would warrant public trust. Thus, we hope to have outlined reasonable and practical means for achieving these worthy goals.

REFERENCES

Atiyeh, C. (2019, November 19). Uber reportedly removed critical auto-braking system on self-driving test car. *Car and Driver*. https://www.caranddriver.com/news/a29834180/uber-auto-braking-removed-accident/

Beauchamp, T. L., & Childress, J. F. (2008). *Principles of biomedical ethics* (6th ed.). Oxford University Press.

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the New Jim Code*. Policy Press.

Blanco, S. (2023, June 13). Report: Tesla Autopilot involved in 736 crashes since 2019. *Car and Driver.* https://www.caranddriver.com/news/a44185487/report-tesla-autopilot-crashes-since-2019/

Bozdag, E. & van den Hoven, J. (2015). Breaking the filter bubble: Democracy and design. *Ethics and Information Technology*, *17*, 249-265. https://doi.org/10.1007/s10676-015-9380-y

Cadillac. (2023). *Super Cruise*. Retrieved August 2023, from https://www.cadillac.com/ownership/vehicle-technology/super-cruise

Cho, H. L., Danis, M., & Grady, C. (2018). Post-trial responsibilities beyond post-trial access. *The Lancet*, *391*(10129), 1478-1479. https://doi.org/10.1016%2FS0140-6736(18)30761-X

Council for International Organizations of Medical Sciences. (2016). *International ethical guidelines for health-related research involving humans* (4th Ed.). Geneva.

Da Silva, M., Flood, C. M., & Herder, M. (2022). Regulation of health-related artificial intelligence in medical devices: The Canadian story. *UBC Law Review*, *55*(3), 635-682. https://commons.allard.ubc.ca/ubclawreview/vol55/iss3/2

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor.* St. Martin's Press.

Fardouly, J., Willburger, B. K., & Vartanian, L. R. (2018). Instagram use and young women's body image concerns and self-objectification: Testing mediational pathways. *New Media & Society*, *20*(4), 1380 – 1395. https://doi.org/10.1177/1461444817694499

Family Smoking Prevention and Tobacco Control Act, United States Code 2009, Title 5, PL 111-31 [HR 1256]. https://www.fda.gov/tobacco-products/rules-regulations-and-guidance/family-smoking-prevention-and-tobacco-control-act-table-contents

Federal Food, Drug & Cosmetic Act, United States Code 2022, Title 21, as amended by P.L. 117–328. https://www.govinfo.gov/content/pkg/COMPS-973/pdf/COMPS-973.pdf

Food and Drug Administration. (2019). *General wellness: Policy for low risk devices*. *Guidance for industry and Food and Drug Administration staff*. https://www.fda.gov/media/90652/download

Food and Drug Regulations, CRC, c 870, 2023. https://laws-lois.justice.gc.ca/PDF/C.R.C._c._870.pdf

Frenkel, S., Confessore, N., Kang, C., Rosenberg, M, & Nicas, J. (2018, November 14). Delay, deny and deflect: How Facebook's leaders fought through crisis. *New York Times*. https://www.nytimes.com/2018/11/14/technology/facebook-data-russia-election-racism.html

Ganz, A. (2018, August 8). IIHS faults Uber for deactivating Volvo's automatic emergency braking in fatal crash. *ABC7Amarillo.com.* Retrieved August 2023, from https://abc7amarillo.com/news/auto-matters/iihs-faults-uber-for-deactivating-volvos-automatic-emergency-braking-in-fatal-crash

Gaspar, J., & Carney, C. (2019). The effect of partial automation on driver attention: A naturalistic driving study. *Human Factors*, *61*(8), 1261-1276. https://doi.org/10.1177/0018720819836310

Gayle, D. (2021, September 14). Facebook aware of Instagram's harmful effect on teenage girls, leak reveals: Social media firm reportedly kept own research secret that suggests app worsens body image issues. *The Guardian.* https://www.theguardian.com/technology/2021/sep/14/facebook-aware-instagram-harmful-effect-teenage-girls-leak-reveals

Holland, G., & Tiggemann, M. (2016). A systemic review of the impact of the use of social networking sites on body image and disordered eating outcomes. *Body Image*, *17*, 100-110. https://doi.org/10.1016/j.bodyim.2016.02.008

Jaynes, N. (2016, July 9). Tesla is the only carmaker beta testing 'autopilot' tech, and that's a problem. *Mashable*. https://mashable.com/article/tesla-beta-testing-autopilot-on-public

Kelly, Y., Silanawala, A., Booker, C., & Slacker, A. (2018). Social media use and adolescent mental health: Findings from the UK millennium cohort study. *eClinical Medicine*, *6*, 59-68. https://doi.org/10.1016/j.eclinm.2018.12.005

Knight, John C. (2002). Safety critical systems: Challenges and directions. *ICSE'02*. https://doi.org/10.1145/581339.581406

Lambert, F. (2016). A fatal Tesla Autopilot accident prompts an evaluation by NHTSA. *Electrek*. https://electrek.co/2016/06/30/tesla-autopilot-fata-crash-nhtsa-investigation/

Llaneras, R. E., Salinger, J., & Green, C. A. (2013). Human factors issues associated with limited ability autonomous driving systems: Drivers' allocation of visual attention to the forward roadway. *Driving Assessment Conference*, *7*(2013), 92-98. https://core.ac.uk/download/pdf/129643921.pdf

Nuremberg Code. (1947). *Trials of war criminals before the Nuremberg military tribunals under control council law, 10*, 181-2.

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism.* New York University Press.

O'Neil, C. (2016). *Weapons of math destruction: How Big Data increases inequality and threatens democracy.* Crown Publishing.

PCMag. (2023). Beta Version. *PCMag Encyclopedia.* Retrieved August 2023, from https://www.pcmag.com/encyclopedia/term/beta-version

Rogerson, S. (2017). Is professional practice at risk following the Volkswagen and Tesla revelations? Software engineering under scrutiny. *ACM Computers & Society*, *3*(47), 25-38. http://dx.doi.org/10.1145/3144592.3144596

Salem, N.F., Le Page, S., Millar, J., Junietz, P., Nolte, M., Graubohm, R., & Maurer, M. (forthcoming 2023). Safety and risk – Why their definitions matter. In, *Handbook on Assisted and Automated Driving*, ser. ATZ/MTZ-Fachbuch, (4th ed.). Springer Vieweg.

Saurwein, F., & Spencer-Smith, C. (2021). Automated trouble: The role of algorithmic selection in harms on social media platforms. *Media and Communication*, *9*(4), 222-233. https://doi.org/10.17645/mac.v9i4.4062

Schwär, H. (2021, August 11). How Instagram and Facebook are intentionally designed to mimic addictive painkillers. *Business Insider.* https://www.businessinsider.com/facebook-has-been-deliberately-designed-to-mimic-addictive-painkillers-2018-12

Shepardson, D. (2023, July 28). Backup driver in 2018 Uber self-driving crash pleads guilty. *Reuters.* https://www.reuters.com/business/autos-transportation/backup-driver-2018-uber-self-driving-crash-pleads-guilty-2023-07-28/

Szilagyi, K., Millar, J., Moon, AJ., Rismani, S. (2023). "Driving into the Loop: Mapping Automation Bias and Liability Issues for Advanced Driver Assistance Systems." *Digital Society 66*. DOI: 10.1007/s44206-023-00066-y

Tesla. (2023a). *Model 3 owner's manual*. Retrieved August 2023, from https://www.tesla.com/ownersmanual/model3/en_us/GUID-2CB60804-9CEA-4F4B-8B04-09B991368DC5.html

Tesla. (2023b). Model S owner's manual. Retrieved August 2023, from *https://www.tesla.com/ownersmanual/2012_2020_models/en_us/GUID-50331432-B914-400D-B93D-556EAD66FD0B.html*

Tiggemann, M., & Zaccardo, M. (2015). 'Exercise to be fit, not skinny': The effect of fitspiration imagery on women's body image. *Body Image*, *15*, 61-67. https://doi.org/10.1016/j.bodyim.2015.06.003

Tobacco and Vaping Products Act, SC 1997, c 13. https://laws-lois.justice.gc.ca/eng/acts/T-11.5/

Wakabayashi, D. (2018, May 24). Emergency braking was disabled when self-driving Uber killed woman, Report Says. *New York Times.* https://www.nytimes.com/2018/05/24/technology/uber-autonomous-car-ntsb-investigation.html#:~:text=The%20Uber%20car%2C%20a%20Volvo,%2C%E2%80%9D%20according%20to%20the%20report.

Wells, G., Horwitz, J., & Seetharaman, D. (2021, September 14). "Facebook Knows Instagram Is Toxic for Teen Girls, Company Documents Show." *The Wall Street Journal.* https://www.wsj.com/articles/facebook-knows-instagram-is-toxic-for-teen-girls-company-documents-show-11631620739

World Health Organization. (2016). Cannabis: The health and social effects of nonmedical cannabis use. https://www.who.int/teams/mental-health-and-substance-use/alcohol-drugs-and-addictive-behaviours/drugs-psychoactive/cannabis&publication=9789241510240

World Health Organization. (2018). Alcohol. https://www.who.int/news-room/fact-sheets/detail/alcohol

World Health Organization. (2021a). Tobacco. https://www.who.int/news-room/fact-sheets/detail/tobacco

World Health Organization. (2021b). Opioid overdose. https://www.who.int/news-room/fact-sheets/detail/opioid-overdose

World Health Organization. (n.d.) Guidelines: Norms and standards for pharmaceuticals. https://www.who.int/teams/health-product-and-policy-standards/standards-and-specifications/norms-and-standards-for-pharmaceuticals/guidelines

World Medical Association. (2013). World Medical Association Declaration of Helsinki: Ethical principles for medical research involving human subjects. Jama, 310(20), pp. 2191-2194. https://doi.org/10.1001/jama.2013.281053