CAN ANDROIDS PLEAD AUTOMATISM? A REVIEW OF *WHEN ROBOTS KILL: ARTIFICIAL INTELLIGENCE UNDER THE CRIMINAL LAW* BY GABRIEL HALLEVY

RACHEL CHARNEY[*]

Humans have long feared the darker side of artificial intelligence (AI). Mutiny, sabotage, and xenocide are themes that science fiction has frequently explored, kindling what Isaac Asimov called the Frankenstein Complex, or the fear of mechanical people.[1] Over time, machines have become both increasingly pervasive and intelligent, allowing us to rely on them to perform a broad range of functions. But does our comfort in allowing machines to clean our homes or beat us in a game of chess mean that we are willing to go so far as to treat humans and machines as similar entities under the criminal law? In *When Robots Kill: Artificial Intelligence Under Criminal Law*, Gabriel Hallevy argues in favour of applying criminal law to artificial intelligence, contending that it would not require any major theoretical revisions to the current legal system.

The issues that Hallevy deals with were raised many years ago,[2] but by expanding on articles that he has previously written Hallevy is the first to set out how the current criminal law framework could be applied to AI.[3] Hallevy begins his book by examining the elusive quest for *machina sapiens*. He explains that although we may never create AI that fully imitates the human mind, criminal liability can still be imposed on machines that are acting under their own agency

[*] Hon BSc (McMaster University), 2nd Year Candidate, JD (University of Toronto). The author would like to thank Sean Gill and the editors of the *University of Toronto, Faculty of Law Review* for their commentary and assistance.

[1] See e.g. *The Terminator*, 1984, DVD (Santa Monica, Cal: MGM Home Entertainment, 2004); *Battlestar Galactica*, 2004, DVD (Universal City, Cal: Universal, 2005); *The Matrix*, 1999, DVD (Warner Bros., 1999); *2001: A Space Odyssey*, 1968, DVD (Burbank, Cal: Warner Home Video, 1999); *Blade Runner*, 1982, DVD (Warner Bros., 1999); See also Lee McCauley, The Frankenstein Complex and Asimov's Three Laws (2007), online: Association for the Advancement of Artificial Intelligence <http://www.aaai.org/Papers/Workshops/2007/WS-07-07/WS07-07-003.pdf>.

[2] See e.g. Raymond S. August, "Turning the Computer Into a Criminal" (1983) 10 Barrister 12; Phil McNally and Sohail Inayatullah, "The rights of robots: Technology, culture and law in the 21st century" (1988) 20:2 Futures 119.

[3] Gabriel Hallevy, "'I, Robot – I, Criminal' – When Science Fiction Becomes Reality: Legal Liability of AI Robots committing Criminal Offenses" (2010) 22 Syracuse Science & Technology L Report 1; Gabriel Hallevy, "The Criminal Liability of Artificial Intelligence Entities – From Science Fiction to Legal Social Control" (2010) 4:2 Akron Intellectual Property J 171; Gabriel Hallevy, "Unmanned Vehicles: Subordination to Criminal Law under the Modern Concept of Criminal Liability" (2011) 21:2 J L Info & Sci 200. More broadly, previous literature has discussed the legal agency of non-human agents and when non-human agents can be awarded legal personhood without specifically examining criminal law (see Sapir Chopra & Laurence F. White, *A Legal Theory for Autonomous Artificial Agents*, (Ann Arbor: The University of Michigan Press, 2011)).

and not merely being used as tools.[4] Hallevy does not provide a clear summary of the minimum capabilities required for AI to achieve agency. Instead, Hallevy avoids the issue by categorizing AI that have the agency to be found criminally liable as 'strong AI', leaving the the reader to piece together the definition of this term.[5]

Hallevy describes in great detail how strong AI can incur criminal liability by meeting both the physical and mental requirements of subjective mens rea offenses, negligence based offenses, and strict liability offenses. He states that while AI of varying competencies can commit the *actus reus*, only more advanced AI will have the processing capacity necessary for the awareness and volition elements that comprise subjective *mens rea*.[6] Negligence and strict liability offences, which do not require subjective *mens rea*, still require the offender to be capable of forming awareness.[7]

Hallevy discusses his framework as though it can be applied to current technology, but modern AI is simply not sufficiently advanced to meet the legal standard of awareness and volition that Hallevy describes.[8] Hallevy draws upon examples of modern day robots but attributes qualities to these robots that are beyond present-day AI capabilities. For example, Hallevy misrepresents medical diagnostic robots by portraying them as commonplace and possessing the awareness required for criminal liability.[9] In reality, the prospect of an AI system that is an accurate bedside diagnostician for a range of ailments is still years away.[10] Similarly, Hallevy distorts the capabilities of robot prison guards in South Korea by suggesting that the guards might face dilemmas that involve prisoner interaction.[11] However, these robot prison guards do not have the capacity to physically interact with prisoners, since they are little more than a camera and mouthpiece for the human prison guards.[12] While our criminal law framework

[4] Gabriel Hallevy, *When Robots Kill: Artificial Intelligence Under Criminal Law* (Boston: Northeastern University Press, 2013) at 74; But see Neil M. Richards and William Smart, How Should the Law Think About Robots? (May 10, 2013), online: SSRN <ssrn.com/abstract=2263363> (the "projection of human attributes is dangerous when trying to design legislation for robots. Robots are, and for many years will remain, tools" at 20).

[5] *Hallevy, supra* note 4 at 6, 56, 64 and 130.

[6] *Ibid* at 40, 49 and 56.

[7] *Ibid* at 87, 109.

[8] See Richards, *supra* note 4 at 6-13.

[9] *Hallevy, supra* note 4 at 84, 95 (although Hallevy never explicitly states that today's medical diagnostic robots would have the awareness required for criminal liability, he implies this when he provides an example that involves a contemporary medical diagnostic robot and uses this example to conclude that as long as AI have met the requirements of a subjective mens rea offense they can be held liable for a negligence based offense).

[10] Jonathan Cohn, "The Robot Will See You Now", *The Atlantic* (20 February 2013), online: The Atlantic Monthly Group <http://www.theatlantic.com >.

[11] *Hallevy, supra* note 4 at 120.

[12] Lena Kim, "Meet South Korea's New Robotic Prison Guards", *Digital Trends* (21 April 2012), online: <http://www.digitaltrends.com/cool-tech/meet-south-koreas-new-robotic-prison-guards/>.

may one day be fully applicable to AI, the current capabilities of AI systems do not reach the legal standard of awareness and volition that our criminal law requires.

Having already set out the framework for holding AI liable for subjective mens rea offenses, negligence offenses, and strict liability offenses, Hallevy next addresses the applicability of general defenses to AI criminal liability. Hallevy provides examples of situations in which each of the defenses he covers (infancy, loss of self control, insanity, intoxication, mistake of fact, mistake of law, and substantive immunity) could be appropriately applied to AI.[13] He also gives examples of situations in which AI could successfully use the justifications of self defence, necessity, duress, following a superior's orders, and *de minimis*.[14] Additionally, Hallevy observes that inducting AI into the criminal law framework does not limit the liability of humans (programmers, users, etc.) who may have been involved in a crime perpetrated by AI. Humans can still be charged as a party to an offense committed by AI, or be found criminally negligent.[15] Hallevy is thorough in each of these sections, and introduces new concepts at a suitable pace.

Hallevy discusses the theory behind sentencing AI and the elements that a judge should consider during sentencing, arguing that punishment can serve a similar purpose for AI as it does for humans. Since machines cannot experience suffering, Hallevy rejects retribution and deterrence as viable justifications for punishment.[16] Yet he does see value in incapacitation and rehabilitation. Incapacitation, such as shutting down or reprogramming the AI, is useful in preventing AI from committing further offenses, and rehabilitation, which involves changing AI's behaviour through machine learning, can allow the AI to make better decisions in the future.[17] While Hallevy mentions the pros and cons of reprogramming and rehabilitation,[18] a thorough discussion on whether an AI's experience is truly lost when it is reprogrammed would have strengthened his argument.

Last, Hallevy discusses possible punishments that AI can receive. Although Hallevy attempts to demonstrate that human punishments are easily applied to AI, a more detailed examination shows errors in his analogies. He argues that capital punishment, imprisonment, and probation can be applied to AI by, respectively, shutting down the AI system, restricting the AI's activities or treating the AI while allowing it to continue its routine activities under court supervision.[19] However, Hallevy's arguments that AI can be punished through public service and fines are unconvincing. On the topic of sentencing through public service, Hallevy presents an example of a medical diagnostic robot. He explains that society could punish

[13] *Hallevy, supra* note 4 at 120-39.

[14] *Ibid* at 140-54.

[15] *Ibid* at 81-82.

[16] *Ibid* at 158-59.

[17] *Ibid* at 160-61.

[18] *Ibid* at 97.

[19] *Ibid* at 165-66, 169 and 171 (treatment for the AI while it is under probation can include intervening in the machine learning process, or upgrading the AI's hardware).

the robot for a negligent diagnosis in a private clinic by forcing it to do a supervised community service placement diagnosing patients in a public hospital.[20] Hallevy ignores that until the robot has been rehabilitated this robot's diagnoses cannot be trusted and through public service the robot might bring harm to the community. Although he is explicit in saying that the robot will be supervised during its public service to prevent further offenses, it is unclear how the supervision would take place, and so it remains possible that further negligence may not be discovered until the harm has already been done. In some cases it may be possible to find an alternative way for an AI system to benefit the community, but unlike humans who can be sentenced to pick up trash after driving over the speed limit, an AI system capable of driving and speeding is not likely to have the capability to pick up trash. As such, the issue of whether the punishment of public service can be applied to AI is a much more nuanced and complicated question than Hallevy indicates.

Furthermore, when Hallevy discusses punishment of AI systems through the use of a fine, he rightly notes that "the main difficulty is that AI systems possess no money or other property of their own".[21] He then suggests that rather than pay a monetary fine, "the [AI] system can pay by using the only currency it possesses: work hours".[22] Modern day law does not recognize AI systems as anything more than property. As such, all means of production of AI systems including its work hours, are owned not by the AI system itself, but by its owner. Consequently, if an AI system was fined through public service then its potentially innocent owners would be unduly punished by having their property confiscated. Hallevy further argues that the purpose of fining an AI system through work hours is not rehabilitative but incapacitive, since additional work hours would allow the robot "less free time to commit offences".[23] This assumes that robots preferentially offend during their 'free time', a proposition unsupported by evidence or logic.

Ultimately, *When Robots Kill* contains original and interesting theoretical ideas, but Hallevy's arguments remain more in the realm of science fiction than practical legal analysis. Although current technology is not sufficiently advanced to accommodate Hallevy's framework, it is entirely possible that emerging technology will allow it to be implemented in the future. While some of his arguments and analogies in the latter half of the book could be more refined, Hallevy's analysis is comprehensive, and leaves out only the question of whether society's resources are truly best spent on AI system rehabilitation. In this book, Hallevy has provided a good starting point for the conversation of whether criminal liability could be applied to AI within our current criminal law framework.

[20] *Ibid* at 172.

[21] *Ibid* at 174.

[22] *Ibid.*

[23] *Ibid.*

NUMBER 2 SPRING 2015 VOLUME 73 NUMÉRO 2 PRINTEMPS 2015

University of Toronto Faculty of Law Review

Revue de droit de l'Université de Toronto

The *University of Toronto Faculty of Law Review* gratefully acknowledges the generous support of its sponsors.

PATRON

JSD Tory Fund

PRIZES AND FELLOWSHIPS[†]

JSD Tory Fellowships are awarded to promising articles with an aim toward eventual publication in the *Law Review*. Laura McGee is this year's recipient of the Torys Fellowship.

The Bill Scadding Essay Prize is awarded for the best article pertaining to family law.

[†] *Members of the Senior Editorial Board of the University of Toronto Faculty of Law Review are not eligible for prizes or awards during their tenure.*

Printed in Canada by Thistle Printing Limited.

Library of Congress Card Number: 54-40057
ISSN: 0381-1638

Subscriptions
The Carswell Company Limited
2075 Kennedy Road
Toronto, Ontario, Canada, M1T 3V4
www.carswell.com

Subscription Rate: $32.00 per issue
Frequency: Two issues per academic year

Advertising Rates
Business Manager
University of Toronto Faculty of Law Review
84 Queen's Park
Toronto, Ontario, Canada, M5S 2C5
E-mail: law.review@utoronto.ca

All Other Communications
The Editors
University of Toronto Faculty of Law Review
84 Queen's Park
Toronto, Ontario, Canada, M5S 2C5
E-mail: law.review@utoronto.ca

Microfiche: A complete set of issues of the *Law Review* is available on microfiche. For more information, please contact William S Hein & Co, Inc, 2350 North Forest Rd., Getzville, NY 14068, USA

Indexing: The *Law Review* is indexed to the Wilson Index to Periodical Literature, LegalTrac, Index to Canadian Periodical Literature, and Scott's Index to Canadian Legal Periodicial Literature.

Online Databases: Full text articles are available through Lexis-Nexis, Quicklaw, William S Hein & Co, Inc, Westlaw Canada, EBSCO, and Cengage Learning. For a full list, please visit our website at www.utflr.org.

Citation: (2015) 73:2 UT Fac L Rev