Τεχνητή νοημοσύνη, αξιόποινες πράξεις, ποινική ευθύνη

Μανώλης Μελισσάρης*

Διδάκτωρ Φιλοσοφίας του Δικαίου, συγγραφέας

Κατά τη λειτουργία των μηχανών τεχνητής νοημοσύνης (ΤΝ) προκαλούνται συχνά βλάβες ή κίνδυνοι βλάβης.[1] Το 2015, ένα σταθερό ρομπότ σε εργοστάσιο της Volkswagen άρπαξε και συνέθλιψε πάνω σε μια μεταλλική πλάκα τον εργάτη που το εγκαθιστούσε, με αποτέλεσμα τον θάνατο του εργάτη.[2] Τα λογισμικά αναγνώρισης προσώπου οδηγούν τακτικά σε άδικες και μεροληπτικές συλλήψεις.[3] Το ChatGPT αναγνώρισε εσφαλμένα έναν καθηγητή νομικής ως δράστη σεξουαλικής παρενόχλησης.[4] Ένας αλγόριθμος με την ονομασία Random Darknet Shopper, που δημιουργήθηκε αλλά δεν ελεγχόταν από δύο καλλιτέχνες, «αγόρασε» παράνομα ναρκωτικά στο διαδίκτυο.[5] Η ΤΝ παραγωγής εικόνων «δημιουργεί» εικόνες κακοποίησης παιδιών σε ανησυχητικά μεγάλη κλίμακα.[6]

Γενικά μιλώντας, οι μηχανές ΤΝ εμφανίζουν ικανότητες, όπως ο συλλογισμός, η μάθηση, η μνήμη και ο σχεδιασμός, που συνήθως συνδέονται αποκλειστικά με τον άνθρωπο. Στην πραγματικότητα, ακριβώς βάσει της ομοιότητας της δραστηριότητάς τους με τη δραστηριότητα των ανθρώπινων δρώντων, αντιπαραβάλλονται συνήθως με τις άλλες, μη ευφυείς μηχανές.

Οι μηχανές ΤΝ πρώτης γενιάς μπορούσαν να επεξεργάζονται έναν πολύ μικρό όγκο δεδομένων και να εκτελούν μάλλον περιορισμένες εργασίες. Μπορούσαν να ολοκληρώνουν έναν τεράστιο αριθμό υπολογισμών πολύ γρήγορα. Οι προγραμματιστές τροφοδοτούσαν τις μηχανές με δεδομένα *και* με πολύ απλούς και σαφείς κανόνες σε μεγάλη κλίμακα. Αυτό έδινε στις μηχανές την ικανότητα να γνωρίζουν σε κάθε περίπτωση τι έπρεπε να «κάνουν» προκειμένου να εκτελούν μια λειτουργία τόσο επιδέξια ώστε συχνά να ξεπερνούν τους ανθρώπους, νικώντας ακόμα και grandmasters στο σκάκι.

---

* Ευχαριστώ τον Δρα Απόστολο Γεωργιάδη για τις πληροφορίες που μου παρείχε αναφορικά με τα τεχνητά νευρωνικά δίκτυα. Είμαι, επίσης, ευγνώμων στον Γιώργο Καράμπελα για τη θαυμάσια απόδοση του κειμένου στα ελληνικά.

1. Οι ενδιαφερόμενοι αναγνώστες μπορούν να παρακολουθήσουν μια αναφορά περιστατικών σχετικών με την ΤΝ που ενημερώνεται τακτικά, στο https://incidentdatabase.ai/.
2. https://www.theguardian.com/world/2015/jul/02/robot-kills-worker-at-volkswagen-plant-in-germany
3. https://theconversation.com/ai-technologies-like-police-facial-recognition-discriminate-against-people-of-colour-143227
4. https://www.washingtonpost.com/technology/2023/04/05/chatgpt-lies/.
5. https://www.bbc.com/future/article/20150721-my-robot-bought-illegal-drugs
6. https://www.theguardian.com/technology/2023/oct/25/ai-created-child-sexual-abuse-images-threaten-overwhelm-internet

Τελικά, θεωρήθηκε ότι αυτό το μοντέλο ήταν πολύ περιορισμένο, κυρίως επειδή απαιτούσε τεράστιο όγκο προγραμματισμού προκειμένου να συνταχθούν εξαντλητικά οι κανόνες που έπρεπε να ακολουθούν οι μηχανές.

Κι εδώ μπαίνει η τεχνολογία των νευρωνικών δικτύων.

Τα πρόσφατα, προηγμένα ρομπότ λειτουργούν με διασυνδεδεμένα τεχνητά νευρωνικά δίκτυα (εφεξής ΤΝΔ), τα οποία είναι μαθηματικές συναρτήσεις φτιαγμένες με πρότυπο τον ανθρώπινο εγκέφαλο. Η τεχνολογία ήταν θεωρητικά γνωστή από καιρό, αλλά η ανάπτυξή της κατέστη δυνατή μόλις τα τελευταία είκοσι χρόνια με την αύξηση της υπολογιστικής ισχύος και τη μείωση του κόστους λειτουργίας. Οι προγραμματιστές δεν παρέχουν πλέον τους κανόνες. Παρέχουν μόνο τα δεδομένα και η μηχανή τα επεξεργάζεται μέσω των ΤΝΔ. Η προηγμένη ΤΝ εμφανίζει τις αναδραστικές δεξιότητες των προκατόχων της, αλλά είναι επίσης σε θέση να κατανοεί ακόμα και ατελείς πληροφορίες και να αναπτύσσει περαιτέρω πρακτικές ικανότητες επίλυσης προβλημάτων με βάση τη γνώση που συσσωρεύει κατά τη διαδικασία.

Το κρίσιμο είναι ότι, σε αντίθεση με τις προγενέστερες μηχανές, αν και γνωρίζουμε τι είδους εργασία θα εκτελέσουν οι μηχανές, δεν έχουμε ιδέα *πώς* θα την εκτελέσουν, τουλάχιστον όχι ακόμα. Μαθαίνουν με το παράδειγμα και την εμπειρία, όχι με την τήρηση κανόνων. Αυτό σημαίνει ότι είναι αυτόνομες και, πολύ σημαντικό, αδιαφανείς και απρόβλεπτες ακόμα και για τους ίδιους τους προγραμματιστές τους. Είναι, όπως το θέλει η έκφραση, «μαύρα κουτιά».

Οι μηχανές αυτές αποτελούν ενδιαφέρουσα πρόκληση όσον αφορά τη νομική ευθύνη γενικά και την ποινική ευθύνη ειδικότερα, και αυτές έχω κατά νου στο παρόν κείμενο.

Πού έγκειται η πρόκληση;

Είναι, χωρίς αμφιβολία, ισχυρή η αίσθηση ότι τα βλαπτικά ή επικίνδυνα περιστατικά, όπως αυτά που περιγράφονται στην εισαγωγή, είναι κάτι περισσότερο από απλά ατυχήματα. Πολλοί το εκφράζουν αυτό με μια αναλογία: αν ένας άνθρωπος σκότωνε ή δυσφημούσε ή αγόραζε παράνομες ουσίες στο διαδίκτυο, θα ήταν ποινικά υπεύθυνος.[7] Κάτι τέτοιο, ωστόσο, συνιστά λήψη του ζητουμένου, καθώς προϋποθέτει ότι έχει διαπραχθεί ένα αδίκημα και ότι, κατά συνέπεια, ισχύουν οι προϋποθέσεις ευθύνης, αν και δεν είναι σαφές πώς ακριβώς. Όμως αυτό ακριβώς είναι το αντικείμενο της έρευνας.

7. Ο Himmelreich, για παράδειγμα, κάνει την αναλογία σε σχέση με αυτόνομες πολεμικές μηχανές. J. Himmelreich, «Responsibility for killer robots», *Ethical Theory and Moral Practice*, 22, 2019, σ. 731-747.

Η αίσθησή μας δεν χρειάζεται να υπερβάλλει τόσο. Ένα ατύχημα είναι κάτι αρκετά απομακρυσμένο από –αν και όχι απαραίτητα εντελώς άσχετο με– τις πράξεις κάποιου. Μπορεί να προκληθεί, ας πούμε, από φυσικά φαινόμενα ή από την αναπόφευκτη φθορά αντικειμένων πέρα από τον άμεσο έλεγχο οποιουδήποτε προσώπου. Στην περίπτωση των μηχανών ΤΝ, όμως, το γεγονός ότι τα ρομπότ προσομοιώνουν την ανθρώπινη συμπεριφορά αρκεί για να αισθανθούμε έντονα ότι η βλάβη δεν προκύπτει τυχαία, ότι δεν θα είχε συμβεί αν δεν είχε προηγηθεί *κάποια* ενέργεια που καθορίζεται από *κάποιου* είδους λήψη αποφάσεων. Αυτό καθιστά απολύτως λογικό να θέσουμε το ερώτημα της ευθύνης, προκειμένου να προσδιορίσουμε αν έχει διαπραχθεί κάποιο αδίκημα.

Ευθύς αμέσως, ωστόσο, σκοντάφτουμε σε εμπόδια. Όπως είδαμε, οι προγραμματιστές ή οι ιδιοκτήτες τους έχουν παραιτηθεί από τον έλεγχο των μηχανών που λειτουργούν με ΤΝΔ. Τροφοδοτούν με δεδομένα το ρομπότ, αλλά ο τρόπος επεξεργασίας τους και τα αποτελέσματα της διαδικασίας καθορίζονται εξ ολοκλήρου από τις «αποφάσεις» και τις «ενέργειες» του ρομπότ. Αυτή η παρέμβαση έχει αρκετή επίδραση στο αποτέλεσμα ώστε να φαίνεται ανεπίτρεπτο να θεωρηθεί ο προγραμματιστής ποινικά υπεύθυνος.

Αυτό μας αφήνει μόνο μία εναλλακτική: να κατηγορήσουμε τη μηχανή. Οι περισσότεροι θα το θεωρούσαν αυτό εξεζητημένο. Οι Brożek και Jakubiec, για παράδειγμα, εκτιμούν ότι οι μηχανές ΤΝ δεν πληρούν ούτε καν εκ του προχείρου τα κριτήρια προκειμένου να θεωρηθούν φορείς νομικής ευθύνης. Υποστηρίζουν ότι το δίκαιο υιοθετεί μια λαϊκή ψυχολογική κατανόηση της ικανότητας πράξης. Νομικώς πράττων είναι αυτός που μπορεί να αποκριθεί στο και να σχετιστεί με το περιβάλλον του με τον ενδεδειγμένο τρόπο, πράγμα που περιλαμβάνει σχέσεις καθήκοντος και ευθύνης. Οι μηχανές, τουλάχιστον προς το παρόν, είναι ανίκανες για κάτι τέτοιο και πρέπει συνεπώς να αποκλειστούν.[8] Οι Brożek και Jakubiec έχουν σε γενικές γραμμές δίκιο, αλλά το επιχείρημα πρέπει να εκλεπτυνθεί και να τεκμηριωθεί με μεγαλύτερη εστίαση. Όπως θα φανεί αργότερα, αυτό θα προσπαθήσω να κάνω στην παρούσα εργασία.

Όμως, ακόμα κι αν τα ρομπότ περάσουν αυτό το πρώτο τεστ καταλληλότητας, οι περισσότεροι φαίνεται να τονίζουν ότι θα αποτύχουν στα επόμενα, καθώς δεν εμφανίζουν, ή κρύβουν πολύ πειστικά, χαρακτηριστικά που απαιτούν οι θεσμοί ευθύνης μας, κυρίως την ικανότητα πρακτικού συλλογισμού που επιτρέπει σε κάποιον να σχηματίζει κρίση. Επομένως, η «τιμωρία» των ρομπότ θα ήταν συγκεχυμένη όσο και επικίνδυνη, διότι θα μπορούσε να υπονομεύσει το κράτος δικαίου και τα ηθικά θεμέλια των θεσμών του ποινικού μας δικαίου.

8. B. Brożek & M. Jakubiec, «On the Legal Responsibility of Autonomous Machines», *Artificial Intelligence and Law*, 25, 2017, σ. 293-304.

Πολλοί θεωρούν απογοητευτικό αδιέξοδο το γεγονός ότι, από τη μια πλευρά, οι μηχανές «κάνουν» πράγματα τα οποία μας προκαλούν αντιδράσεις που κανονικά θα επιφυλάσσαμε για επιλήψιμες συμπεριφορές,[9] ενώ, από την άλλη πλευρά, λόγω της φύσης των ρομπότ και των θεμελιωδών αρχών των θεσμών ευθύνης μας, κανείς δεν μπορεί στ' αλήθεια να θεωρηθεί υπεύθυνος, πόσο μάλλον να λογοδοτήσει. Αυτό το «κενό ευθύνης», όπως το ονόμασε ο Andreas Matthias ήδη το 2004, πρόκειται να διευρυνθεί ακόμα περισσότερο όσο η τεχνολογία εξελίσσεται και συνεχίζει να υπονομεύει τις βασικές μας πρακτικές καταλογισμού ευθύνης και επιβολής τιμωρίας.[10] Η λύση που πρότεινε ο Matthias ήταν να περιοριστεί η ανάπτυξη και η εξάπλωση της Τεχνητής Νοημοσύνης. Σχεδόν είκοσι χρόνια μετά, η έκκλησή του προφανώς έχει μέχρι στιγμής πέσει στο κενό.[11]

Ορισμένοι συμμερίζονται την ανησυχία του Matthias και έχουν αποδεχθεί και αναπτύξει περαιτέρω την ιδέα του κενού ευθύνης.[12] Άλλοι δεν έχουν πειστεί ότι υπάρχει καν τέτοιο κενό.[13]

Τέλος, υπάρχουν και εκείνοι που δεν εγκαταλείπουν την εννοιολογική δυνατότητα απόδοσης ευθύνης στις μηχανές ΤΝ. Εστιάζοντας αποκλειστικά στην ποινική ευθύνη και υποθέτοντας ότι το πρόβλημα σχετίζεται με τη mens rea (υποκειμενική υπόσταση), οι Abbott και Sarch προτείνουν τρεις πιθανούς τρόπους για να βγούμε από το αδιέξοδο.[14] Ο πρώτος είναι να καταλογιστούν στις μηχανές οι νοητικές και βουλητικές καταστάσεις των δημιουργών ή των αρχικών χρηστών τους. Η δεύτερη επιλογή είναι ένα είδος αυστηρής (αντικειμενικής) ευθύνης. Η τρίτη, και πιο ενδιαφέρουσα για τους σκοπούς μας, είναι η επιβολή άμεσης ποινικής ευθύνης στον βαθμό που «η ΤΝ είναι προγραμματισμένη ώστε να είναι σε θέση να λαμβάνει υπόψη

9. Αυτό απηχεί τη θεώρηση του Strawson για την ευθύνη. Βλ. P. F. Strawson, «Freedom and Resentment», στο *Proceedings of the British Academy*, 48, 1962, σ. 1-25.

10. A. Matthias, «The responsibility gap: Ascribing responsibility for the actions of learning automata», *Ethics and Information Technology*, 6(3), 2004, σ. 175-183.

11. Όσο γραφόταν αυτό το κείμενο, διεξάχθηκε μια σύνοδος ηγετικών παραγόντων της πολιτικής, της τεχνολογίας και του επιχειρηματικού τομέα για να συζητηθούν οι κίνδυνοι που θέτει η ανάπτυξη και εξάπλωση της χρήσης της ΤΝ και οι τρόποι ρύθμισής τους (london.theaisummit.com/)

12. Βλ. F. Santoni de Sio & G. Mecacci, «Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address Them», *Philosophy and Technology*, 34, 2021, σ. 1057-1084· J. Danaher, «Robots, law and the retribution gap», *Ethics and Information Technology*, 18(4), 2016, σ. 299-309· R. Sparrow, «Killer robots», *Journal of Applied Philosophy*, 24(1), 2007, σ. 62-77.

13. Βλ. S. Köhler, N. Roughley, H. Sauer, «Technologically blurred accountability», στο C. Ulbert, P. Finkenbusch, E. Sondermann, T. Diebel (επιμ.), *Moral agency and the politics of responsibility*, Routledge 2018, σ. 51-68· D.R. Tigard, «There Is No Techno-Responsibility Gap», *Philosophy & Technology*, 34, 2021, σ. 589-607· F. Hindricks & H. Veluwenkamp, «The risks of autonomous machines: from responsibility gaps to control gaps», *Synthese*, 201, 2023, σ. 21.

14. R. Abbott and A. Sarch, «Punishing Artificial Intelligence: Legal Fiction or Science Fiction», *UC Davis Law Review* 53(1), 2019, σ. 323-384.

τα συμφέροντα των ανθρώπων και να συνυπολογίζει νομικές απαιτήσεις, αλλά καταλήγει να συμπεριφέρεται με τρόπο που δεν συνάδει με την ορθή συνεκτίμηση αυτών των νομικά αναγνωρισμένων συμφερόντων και λόγων», διότι τότε το ρομπότ επιδεικνύει εκείνη την αδιαφορία για τους άλλους που αποτελεί αντικείμενο μομφής από το ποινικό δίκαιο.

Παρομοίως, ο Gabriel Hallevy πιστεύει ότι οι μηχανές ΤΝ είναι γνωστικά εξοπλισμένες προκειμένου να σχηματίζουν τη γνώση που απαιτείται από τους σκοπούς του ποινικού δικαίου. Εφόσον οι πράξεις τους πληρούν επίσης όλες τις σχετικές προϋποθέσεις, υποστηρίζει ο Hallevy, δεν υπάρχουν εννοιολογικά εμπόδια για τη θεμελίωση ποινικής ευθύνης.[15]

Ρομποτικές πράξεις

Έχουμε την τάση να φανταζόμαστε τις μηχανές ΤΝ να οδηγούν, να γράφουν, να κάνουν διαγνώσεις, να προβλέπουν, να αναγνωρίζουν – με άλλα λόγια, να ενεργούν με τον ίδιο τρόπο που ενεργούμε κι εμείς. Είναι όντως έτσι ακριβώς;

Οι περίπου εκατό δισεκατομμύρια νευρώνες στο σώμα μας, που βρίσκονται κυρίως στον εγκέφαλο και στη σπονδυλική στήλη, αλλά και σε περιφερικά γάγγλια, είναι κύτταρα που δέχονται αισθητηριακές εισροές από τον εξωτερικό κόσμο, επικοινωνούν μεταξύ τους και στέλνουν κινητικές εντολές στους μυς μας. Τα ερεθίσματα λαμβάνονται από τους δενδρίτες, που ονομάζονται έτσι επειδή μοιάζουν με κλαδιά δέντρων. Στη συνέχεια, ένα σήμα μεταβιβάζεται στον άξονα, ο οποίος διατρέχει σαν καλώδιο τον νευρώνα. Ο άξονας μεταδίδει ένα ηλεκτρικό φορτίο στη σύναψη, το σημείο σύνδεσης μεταξύ των νευρώνων (κάθε νευρώνας έχει περίπου χίλιες συνάψεις, οι οποίες αθροιζόμενες φτάνουν στο απίθανο νούμερο των εκατό τρισεκατομμυρίων). Αυτό στη συνέχεια μετατρέπεται για λίγο σε χημικό νευροδιαβιβαστή και κατόπιν ξανά σε ηλεκτρικό σήμα στον δενδρίτη ενός άλλου νευρώνα. Τα μηνύματα μεταφέρονται τελικά στους μυς, οι μυϊκές ίνες συσπώνται και το σώμα μας κινείται.
Τα ΤΝΔ λειτουργούν με τον ίδιο τρόπο, μόνο που αυτή τη φορά τα σήματα είναι σε κώδικα ανθρώπινης κατασκευής και οι νευρώνες είναι μαθηματικές συναρτήσεις. Με απλά λόγια, πατώντας το πληκτρολόγιό μας στέλνουμε τάσεις στον υπολογιστή, δηλαδή δυαδικά σήματα 01, τα οποία αποτελούν τα δεδομένα εισόδου. Αυτό ενεργοποιεί τον αλγόριθμο, γίνονται οι απαραίτητοι υπολογισμοί και μεταβιβάζονται από νευρώνα σε νευρώνα στα κρυμμένα επίπεδα, με μερικά μπρος-πίσω, μέχρι να ικανοποιηθεί η μηχανή ότι έχει την ορθή απάντηση ή σειρά ενεργειών στο επίπεδο εξόδου.

15. G. Hallevy, «The Criminal Liability of Artificial Intelligence Entities – From Science Fiction to Legal Social Control», *Akron Intellectual Property Journal*, 4:2, 2010, σ. 171-201.

Για να μπορέσει η μηχανή να παραγάγει ουσιαστικά αποτελέσματα, πρέπει να εκπαιδευτεί. Σε αντίθεση με τις μηχανές παλαιότερης τεχνολογίας, οι οποίες απαιτούσαν λεπτομερείς κανόνες που συντάσσονταν από τον προγραμματιστή και ήταν ανά πάσα στιγμή γνωστοί σε αυτόν, τα ρομπότ νέας γενιάς μαθαίνουν με το παράδειγμα και την εμπειρία. Λόγου χάριν, το ChatGPT τροφοδοτείται με μια μικροσκοπική μπουκιά δεδομένων τη φορά, έτσι ώστε, όταν του παρουσιαστεί την επόμενη φορά η ίδια ακολουθία, να είναι σε θέση να υπολογίσει τι έπεται. Σκεφτείτε ότι τρισεκατομμύρια κείμενα, σχεδόν το σύνολο του περιεχομένου του διαδικτύου, έχουν εισαχθεί στο Chat GPT και θα δείτε πώς καταφέρνει να συνθέτει με επιτυχία, τουλάχιστον εκ πρώτης όψεως, ολοκληρωμένα κείμενα ως απάντηση σε εντολές ή ερωτήσεις.

Έτσι, σε ένα επίπεδο, θα μπορούσε κανείς να υποστηρίξει αρκετά εύλογα ότι δεν υπάρχει καμία διαφορά μεταξύ ρομποτικών και ανθρώπινων πράξεων. Όταν οι υπολογιστές βλέπουν, ας πούμε, μια γάτα και την αναγνωρίζουν ως τέτοια, αναλύουν, με αδιανόητη ταχύτητα, απίστευτη ακρίβεια και, όπως ήδη αναφέραμε, σχεδόν τέλεια αδιαφάνεια, την εικόνα μιας γάτας σε pixels, αντιπαραβάλλουν τα μέρη με τα δεδομένα που έχουν συσσωρεύσει και καταλήγουν στο συμπέρασμα ότι πρόκειται πράγματι για γάτα. Όταν βλέπουμε *εμείς* μια γάτα, το φως προσκρούει στον αμφιβληστροειδή χιτώνα και στη συνέχεια μετατρέπεται σε ηλεκτρικό σήμα, το οποίο μεταδίδεται στον εγκέφαλο για να αναλάβουν δράση οι νευρώνες και τελικά (εκτός από τυχόν δυσλειτουργίες, όλα αυτά συμβαίνουν σε νανοδευτερόλεπτα, γεγονός που κάνει τη διαδικασία να γίνεται αισθητή σαν στιγμιαία) να αναγνωρίσουμε το αντικείμενο ως γάτα.

Σε κάθε περίπτωση, εδώ μάς ενδιαφέρει κάτι πιο συγκεκριμένο, δηλαδή αν μπορούμε να πούμε ότι οι μηχανές ΤΝ ενεργούν *με τρόπο που άπτεται του ποινικού δικαίου*, ότι οι πράξεις τους ικανοποιούν την απαίτηση της actus reus (αντικειμενικής υπόστασης).

Actus reus και η υπόθεση της συνείδησης

Όπως ανέφερα προηγουμένως, ορισμένοι υποθέτουν ότι οι μηχανές ΤΝ όχι μόνο πράττουν αλλά και πράττουν κατά τρόπο καταλογιστό εις ενοχή.[16] Εφόσον πληρούται η απαίτηση της actus reus, το μόνο σημαντικό και καίριο ερώτημα είναι κατά πόσον οι υπολογιστές μπορούν να σχηματίσουν την απαιτούμενη mens rea. Θεωρώ αυτή την άποψη μάλλον βιαστική και τελικά λανθασμένη, διότι υποτιμά τη φύση της actus reus.

16. Συνήθως δε, αναμφίλεκτα. Για παράδειγμα: *«Είναι σχετικά απλό να αποδώσουμε actus reus σε ένα σύστημα ΤΝ. Αν ένα σύστημα αναλαμβάνει μια δράση που καταλήγει σε αξιόποινη πράξη, ή δεν καταφέρνει να αναλάβει δράση ενώ είναι καθήκον του να δράσει, τότε έχει προκύψει η actus reus ενός αδικήματος»*. J.K.C. Kingston, «Artificial Intelligence and Legal Liability», στο *Research and Development in Intelligent Systems XXXIII: Incorporating Applications and Innovations in Intelligent Systems XXIV*, Springer 2016, σ. 269-279.

Οι άνθρωποι θεωρούν ότι έχουν αυτό που έχει επικρατήσει να ονομάζεται συνείδηση. Δεν υπάρχει μεγάλη συναίνεση μεταξύ των επιστημόνων και των φιλοσόφων ως προς το τι είναι η συνείδηση, πόσο μάλλον αν έχει έστω πραγματικό αντίστοιχο. Εξ όσων γνωρίζουμε, μπορεί να είμαστε όλοι συνδεδεμένοι σε έναν υπολογιστή και αυτό που θεωρούμε ως μοναδική ικανότητα που μας διαφοροποιεί από τα άλλα όντα να είναι μια ψευδαίσθηση ή μια απάτη που σκαρώνεται εις βάρος μας. Εν τούτοις, αυτό δεν αλλάζει την *αντίληψή* μας για τα επίπεδα της ύπαρξής μας. Υιοθετώντας την οπτική του πρώτου προσώπου, βιώνουμε και αναπαριστούμε τον κόσμο όπως φαίνεται να είναι και τοποθετούμαστε μέσα σε αυτόν σε μια σχέση αλληλεξάρτησης – οι αλλαγές στον κόσμο μάς επηρεάζουν, και αντιστρόφως. Το κάνουμε αυτό με μια σχετική συνοχή στον χώρο και στον χρόνο, καθώς θυμόμαστε τον εαυτό μας στο παρελθόν και φανταζόμαστε τον εαυτό μας στο μέλλον. Ταυτόχρονα, αποστασιοποιούμαστε από αυτόν, όντας έτσι σε θέση να αναπτύξουμε μια στοχαστική στάση απέναντί του.

Αυτή η υπόθεση της συνείδησης είναι ενσωματωμένη στους θεσμούς μας, με κυριότερο από αυτούς το δίκαιο. Έχει κομβική σημασία ότι στηρίζει τις αντιλήψεις περί νομικής ευθύνης και αντανακλάται στην απαίτηση της πράξης στο ποινικό δίκαιο με δύο τρόπους.

α) Ο πρώτος είναι μάλλον προφανής, αλλά παραδόξως απουσιάζει από τη συζήτηση για την ΤΝ και την ποινική ευθύνη. Όταν το ποινικό δίκαιο απαιτεί από τα υποκείμενά του να ενεργούν (ή να μην ενεργούν, όταν έχουν καθήκον να ενεργήσουν) για να είναι ποινικά υπεύθυνα, απαιτεί να το κάνουν συνειδητά, με αυτεπίγνωση και με βουλητικό έλεγχο των κινήσεών τους. Με άλλα λόγια, αναμένει να έχουν πρόθεση να εκτελέσουν τις κινήσεις που επιφέρουν το απαιτούμενο αποτέλεσμα. Η actus reus, η αντικειμενική όψη του αδικήματος, δεν είναι απαλλαγμένη από ψυχικά στοιχεία, αν και αυτή η υπολειπόμενη υποκειμενικότητα παραμένει κανονιστικά ουδέτερη και πρέπει να διακρίνεται προσεκτικά από την πρόθεση διάπραξης της ενέργειας ως αδικήματος, η οποία σχετίζεται με μια ηθική στάση που πρέπει να έχει αναπτύξει ο κατηγορούμενος.

β) Εκτός από την πρόθεση να μετακινήσει κανείς το σώμα του, τα αδικήματα του ποινικού δικαίου συνήθως απαιτούν επίσης γνώση ή μια πεποίθηση αναφορικά με τις περιστάσεις διάπραξης ενός αδικήματος, με τη διαφορά ότι η γνώση είναι δικαιολογημένη πεποίθηση, ενώ η απλή πεποίθηση είναι καθαρά υποκειμενική. Θα υποστήριζα, ωστόσο, ότι και οι δύο αυτές απαιτήσεις εντάσσονται στο πλαίσιο μιας ευρύτερης ικανότητας την οποία αποδίδει το ποινικό δίκαιο στα συνειδητά υποκείμενά του.

Ακολουθεί ένα παράδειγμα. Ένας από τους τρόπους διάπραξης του αδικήματος της απάτης στην Αγγλία και την Ουαλία είναι η ψευδής παράσταση. Σύμφωνα με τα άρθρα 1 και 2 του

νόμου περί απάτης του 2006, οι προδιαγραφές της actus reus του αδικήματος είναι: η παροχή μιας παράστασης, σε σχέση με ένα γεγονός («είμαι ο κληρονόμος του πλουσιότερου ανθρώπου στον κόσμο»), έναν κανόνα δικαίου (π.χ. «μου χρωστάς Χ χρηματικό ποσό») ή την πνευματική κατάσταση του προσώπου που παρέχει την παράσταση ή ενός τρίτου προσώπου (π.χ. «το αφεντικό σου σε διατάζει να μου δώσεις Χ χρηματικό ποσό»), η οποία γίνεται είτε προφορικά, είτε εγγράφως, είτε μέσω μηχανής (ώστε να καλύπτεται η δόλια χρήση τραπεζικών καρτών και τα παρόμοια). Σημειωτέον ότι δεν απαιτείται η παράσταση να έχει οποιοδήποτε αποτέλεσμα. Ούτε ο αποδέκτης της χρειάζεται να πέσει θύμα της παραπλάνησης, ούτε το πρόσωπο που την κάνει χρειάζεται να αποκομίσει κέρδος ή να προκαλέσει ζημία (αν και η πρόθεση να επιφέρει το ένα ή το άλλο αποτελεί προϋπόθεση της mens rea του αδικήματος).

Για να διαπράξει κάποιος απάτη με ψευδή παράσταση, πρέπει πράγματι να εκτελέσει εκούσια ορισμένες σωματικές κινήσεις – για παράδειγμα, να εκφέρει ή να πληκτρολογήσει τις λέξεις που περιγράφουν μια κατάσταση. Αυτό από μόνο του, ωστόσο, δεν αρκεί. Η εκφορά αναφέρεται σε κάτι έξω από τη σωματική της εκδήλωση και αντλεί το νόημά της από αυτό το απομακρυσμένο επίπεδο. Η ύπαρξη, με τον ένα ή τον άλλο τρόπο, αυτού του επιπέδου είναι επομένως ενσωματωμένη στην πράξη της παράστασης. Δεν πρόκειται εδώ για μια φιλοσοφική άποψη σχετικά με τη φύση και τη λειτουργία της γλώσσας. Ο νόμος δεν ασχολείται με σχολαστικά ζητήματα όπως αν το νόημα είναι κοινωνικά κατασκευασμένο ή εγγενές στον κόσμο. Ενσωματώνει μια κοινή, λαϊκή άποψη, την οποία οι περισσότεροι άνθρωποι αναγνωρίζουν και αποδέχονται ως εύλογη, αναφορικά με το πώς σχετιζόμαστε με τον κόσμο και του αποδίδουμε νόημα.

Αν και το άρθρο 2 του νόμου περί απάτης δεν το αναφέρει ρητά, δεν αρκεί να κάνει ο κατηγορούμενος μια εκφορά που ισοδυναμεί με παράσταση. Η απάτη είναι αδίκημα παραπλάνησης, παρόλο που δεν απαιτείται να παραπλανηθεί πραγματικά κάποιος,[17] και για να συμβεί αυτό χρειάζονται δύο, ένας πομπός και ένας δέκτης. Αυτό προϋποθέτει ότι η δήλωση που ισοδυναμεί με παράσταση πρέπει *να μπορεί να κοινοποιηθεί* σε κάποιον άλλον, πράγμα που με τη σειρά του συνεπάγεται, μεταξύ άλλων, ότι η παράσταση πρέπει να είναι κατανοητή και να γίνεται με αποτελεσματικό τρόπο. Οι εν λόγω προϋποθέσεις εξειδικεύονται περαιτέρω σε σχέση με τον τρόπο λειτουργίας. Για παράδειγμα, η δόλια παράσταση μέσω υπολογιστή προϋποθέτει ότι άλλα, απομακρυσμένα μέρη μπορούν να αποκτήσουν πρόσβαση στην παράσταση με την κατάλληλη τεχνολογία, κ.ο.κ. Ομοίως, η απαίτηση να είναι η παράσταση αναληθής ή παραπλανητική εξαρτάται από την πιθανή ύπαρξη μιας διαφορετικής περιγραφής των γεγονότων καθώς και ενός φόρουμ ή κριτηρίων για να αποφασιστεί ποια εκδοχή των γεγονότων είναι αληθής ή

17. Μάλιστα, αμέσως μόλις ψηφίστηκε ο νόμος, κατακρίθηκε ότι ποινικοποιούσε το ψέμα. Βλ. λ.χ. D. Ormerod, «The Fraud Act 2006 – criminalising lying?», *Criminal Law Review*, 2007, σ. 193.

ακριβής. Και για να γνωρίζει ή να υποπτεύεται ο Δ ότι η παράσταση είναι ή μπορεί να είναι ψευδής, πρέπει να έχει πρόσβαση σε αυτά τα κριτήρια απόφασης.

Το ποινικό δίκαιο της απάτης αναμένει από τον κατηγορούμενο να είναι σε θέση να κατανοήσει αυτές τις βασικές προϋποθέσεις της ψευδούς παράστασης. Αυτό περιλαμβάνει την ικανότητα να γνωρίζει ή να πιστεύει συγκεκριμένες περιστάσεις, αλλά, αν το δούμε ευρύτερα, και αυτό με τη σειρά του εξαρτάται από τη γενικότερη ικανότητα να σχηματίζει κανείς μια εικόνα του κόσμου ως πραγματικότητας και ως δυνατότητας. Μολονότι η ικανότητα να έχει κανείς μια αίσθηση της ευρύτερης εικόνας δεν είναι από μόνη της θέμα ούτε γνώσης, κυρίως επειδή μεγάλο μέρος της αποτελείται από μελλοντικές προβλέψεις, ούτε πίστης, κυρίως επειδή μπορεί να περιλαμβάνει ανταγωνιστικές και ασυμβίβαστες εναλλακτικές, τις οποίες δεν μπορεί κανείς να πιστεύει ταυτόχρονα, αποτελεί προϋπόθεση για να γνωρίζει ή να πιστεύει τις ιδιαίτερες συνθήκες γύρω από τη διάπραξη του αδικήματος. Θα μπορούσαμε να το ονομάσουμε αυτό ζήτημα *φαντασίας*, έναν ορίζοντα επίγνωσης της κατάστασης του κόσμου με το γενικά αποδεκτό κοινωνικό του νόημα. [18] [19]

Εδώ απλώς σκιαγραφώ βιαστικά κάτι πολύπλοκο και ενδεχομένως αμφιλεγόμενο στις λεπτομέρειές του. Είμαι, ωστόσο, βέβαιος ότι η βασική ιδέα συνάδει με μια ευρέως διαδεδομένη κατανόηση του δικαίου και ότι είναι, ως εκ τούτου, αδιαμφισβήτητη. Αν ισχύει αυτό, τότε αρκεί για να υποστηρίξουμε τα εξής. Όπως η δυνατότητα βουλητικού ελέγχου των κινήσεών μας είναι απόρροια της υπόθεσης της συνείδησης στην οποία στηρίζεται το δίκαιο, το ίδιο είναι και η ικανότητα της φαντασίας, η ικανότητα να τοποθετούμε τις πράξεις μας σε ένα ευρύτερο πλαίσιο, για το οποίο σχηματίζουμε μια νοερή εικόνα. Έπεται ότι όποιος δεν έχει αυτή τη σχετιζόμενη με τη συνείδηση ικανότητα της φαντασίας είναι ανίκανος να ενεργήσει ως υποκείμενο του ποινικού δικαίου, όπως και όποιος δεν έχει την ικανότητα να κάνει εκούσιες κινήσεις δεν μπορεί να διαπράξει ποινικό αδίκημα.

Αυτό δεν σημαίνει ότι το υποκείμενο του ποινικού δικαίου πρέπει να ασκεί *σωστά* ή με *ακρίβεια* την ικανότητα της φαντασίας, την οποία ο νόμος θεωρεί ότι έχει, για να θεωρηθεί η πράξη του αξιόποινη. Σκεφτείτε τη θεωρία της πραγματικής αδυναμίας. Στο δίκαιο της Αγγλίας και της Ουαλίας, και αναμφίβολα και σε άλλα δικαιικά συστήματα, η πλάνη ως προς τις πραγμα-

18. Φαίνεται να υπάρχει ολοένα μεγαλύτερο ενδιαφέρον για τη φαντασία στο δίκαιο. Βλ. για παράδειγμα, M. Del Mar, *Artefacts of Legal Inquiry: The Value of Imagination in Adjudication*, Bloomsbury 2020, που ασχολείται με τη φαντασία στις δικαστικές αποφάσεις.

19. Θα υποστήριζα ότι η ικανότητα της φαντασίας υπεισέρχεται ακόμα και στα αδικήματα, όπου φαίνεται ότι το μόνο που απαιτείται είναι ο εκούσιος έλεγχος των κινήσεων του δράστη, αν και το επιχείρημα αυτό θα πρέπει να διατυπωθεί σε διαφορετικά συμφραζόμενα.

τικές περιστάσεις της πράξης δεν αποτελεί υπεράσπιση. Για την απόδοση ποινικής ευθύνης, τα γεγονότα λαμβάνονται έτσι όπως τα πίστευε ο κατηγορούμενος. Για να επανέλθουμε στο τρέχον παράδειγμά μας, αν, ας πούμε, ο κατηγορούμενος κάνει την απατηλή παράσταση στέλνοντας τηλεπαθητικά σήματα ή μιλώντας σε μια γλώσσα που ο κατηγορούμενος έχει επινοήσει εξ ολοκλήρου και την οποία δεν μιλάει κανείς άλλος, ο κατηγορούμενος εξακολουθεί να είναι υπεύθυνος για απόπειρα απάτης. Όμως το σημαντικό έγκειται στο ότι προϋπόθεση της ευθύνης είναι ο κατηγορούμενος να έχει τη συνειδητή γνωστική ικανότητα να φαντάζεται την πράξη του μέσα σε ένα πλαίσιο το οποίο περιλαμβάνει τις συνέπειές της και τις αντιδράσεις των άλλων σε αυτή.

Θα μπορούσε να υποστηρίξει κανείς ότι η ψυχική κατάσταση που συνοδεύει τη σύννομη δράση μπορεί επίσης να εξηγηθεί μικροσκοπικά με όρους λειτουργίας των νευρωνικών δικτύων, και ότι όροι όπως «πρόθεση» δηλώνουν απλώς τις λειτουργίες των νεύρων και των μυών μας. Ενδεχομένως – ωστόσο η άποψη αυτή δεν συνάδει με την αντίληψη του ποινικού δικαίου για το δρών υποκείμενο. Το δίκαιο υποθέτει ότι η συνειδητή πρόθεση και η φαντασία, αφενός, και η πράξη ως σωματική κίνηση, αφετέρου, είναι δύο διακριτά επίπεδα, ότι το υποκείμενο του ποινικού δικαίου μπορεί πραγματικά να λάβει μια απομακρυσμένη θέση σε σχέση με την ύπαρξή του, η οποία φυσικά περιλαμβάνει τις ίδιες τις κινήσεις του. Αυτό τού επιτρέπει να επιτελεί δύο κρίσιμες λειτουργίες.

Πρώτον, η υπόθεση της συνείδησης, η οποία μεταφράζεται στις απαιτήσεις της εκούσιας βούλησης και της φαντασίας, *αποδίδει την επιβλαβή πράξη στον δράστη με έναν τρόπο που συγκεκριμενοποιεί την ευθύνη*. Σκεφτείτε τα αυτόματα. Ας πούμε ότι το χέρι του Α σηκώνεται στον αέρα και χτυπά τον Β. Στο πρώτο σενάριο, το χέρι του κινείται ακούσια λόγω μυϊκού σπασμού. Αν θεωρήσουμε την πρόθεση να κινηθεί το χέρι ως προϊόν νευρωνικών αλληλεπιδράσεων, τότε τίποτα δεν μας εμποδίζει να πούμε ότι ο Α χτυπάει εκούσια τον Β. Στην καθημερινή ορολογία, μπορεί μάλιστα να το κάνουμε χωρίς να το σκεφτόμαστε, αν και δεν θα προχωρούσαμε στο να κατηγορήσουμε τον Α γι' αυτό. Κάτι τέτοιο θα ήταν ανακριβές από την άποψη του ποινικού δικαίου. Δεν είναι ότι ο Α δεν ευθύνεται, επειδή δεν είχε έλεγχο των πράξεών του: *δεν ενήργησε καθόλου*, διότι δεν θέλησε τις μυϊκές του κινήσεις. Το να κινείται ο Α ως αυτόματο δεν ισοδυναμεί με το να ενεργεί ο Α ως συνειδητό υποκείμενο.

Αυτό καθιστά επίσης δυνατή την απόδοση απομακρυσμένων βλαβών στον πράττοντα. Η παρουσίαση ενός κινδύνου, για παράδειγμα, απαιτεί από τον δρώντα την ικανότητα πρόβλεψης των πιθανών αποτελεσμάτων των πράξεών του, η οποία μπορεί να είναι χαρακτηριστικό μόνο ενός υποκειμένου που έχει μια νοερή αντίληψη της ευρύτερης εικόνας του κόσμου. Δεν θέλω με αυτό να πάρω θέση ως προς το αν είναι σωστό να αποδίδονται στους ανθρώπους ευθύνες

για απομακρυσμένες βλάβες ή για τη δημιουργία κινδύνων. Θέλω απλώς να πω ότι η υπόθεση της συνείδησης αποτελεί αναγκαία προϋπόθεση για να αρχίσει κανείς να εξετάζει την απόδοση ποινικής ευθύνης για απομακρυσμένες βλάβες ή για τη δημιουργία κινδύνων.

Δεύτερον, η υπόθεση της συνείδησης και οι εξειδικεύσεις της βοηθούν στην *εξατομίκευση του δράστη*. Θεωρούμε ότι μια πράξη έχει τελεστεί *από τον κατηγορούμενο*, και όχι από κάποιον άλλον, διότι, και μόνο καθόσον, είχε τον έλεγχο των σχετικών σωματικών του κινήσεων *και* τις ήθελε. Πρόκειται για μια πιο εκλεπτυσμένη εκδήλωση της εξατομικευτικής λειτουργίας της συνείδησης ως υποκειμενικότητας, η οποία στο ποινικό δίκαιο συμπληρώνεται με τις υπόλοιπες απαιτήσεις που αφορούν τη mens rea. Το ποινικό δίκαιο δεν εκλαμβάνει τα υποκείμενά του ως διαδικασίες, αλλά ως κεντρικά ελεγχόμενες δέσμες αλληλεπικαλυπτόμενων και αλληλοαποκρινόμενων σωματικών και ψυχικών χαρακτηριστικών, και υπό αυτή την προϋπόθεση κατανέμει την ευθύνη. Αυτό το υποκείμενο θεωρείται πάντα το ίδιο στον χρόνο.

Τεχνητή νοημοσύνη και actus reus

Εξ όσων γνωρίζουμε και όπως υφίσταται σήμερα η τεχνολογία, οι μηχανές, όσο καλά εκπαιδευμένες κι αν είναι, ανταποκρίνονται σε ηλεκτρικά σήματα που ενεργοποιούν μαθηματικές συναρτήσεις. Ακολουθεί μια σύνθετη, αμφίδρομη αλληλεπίδραση τεχνητών νευρώνων, με αποτέλεσμα, ας πούμε, να κινείται ένα άλλο αντικείμενο, κείμενα να προβλέπουν μελλοντικά γεγονότα ή να διαγιγνώσκουν μια παθολογική κατάσταση, κ.ο.κ. Οι μηχανές είναι τόσο εντυπωσιακά επιδέξιες και, στην πλειονότητά τους τουλάχιστον, τόσο αποτελεσματικές στην εκτέλεση των καθηκόντων τους ώστε τείνουμε να πιστεύουμε ότι ενεργούν ακριβώς όπως εμείς. Θεωρούμε κατά βάση ότι οι πράξεις τους είναι όπως οι δικές μας, δηλαδή ότι παρακινούνται και ελέγχονται από κάποιο κέντρο που απέχει από την καθαρά σωματική μας ύπαρξη – το κέντρο που ονομάζουμε συνείδηση. Κατ' επέκταση, δεχόμαστε επίσης ότι οι μηχανές ΤΝ ενεργούν με ποινικά υπεύθυνο τρόπο, μια εντύπωση που αντανακλάται στις γλωσσικές μας πρακτικές για τις προηγμένες μηχανές γενικά και για όσες σχετίζονται με το ποινικό δίκαιο ειδικότερα. Τείνουμε να λέμε ότι το ρομπότ στη Volkswagen «σκότωσε» τον εργάτη ή ότι το ChatGPT «δυσφήμησε» τον καθηγητή νομικής επειδή υποθέτουμε ότι αυτές οι πράξεις ανήκουν στη μηχανή όπως ακριβώς θα ανήκαν σε ένα ανθρώπινο ον.

Είναι απλή εντύπωση αυτό ή κάτι παραπάνω; Ο Jaap Hage προσφέρει ένα πιο εκλεπτυσμένο επιχείρημα υπέρ της δυνατότητας να θεωρηθούν οι μηχανές ΤΝ κανονικά δρώντα υποκείμενα

του ποινικού δικαίου.[20] Δεδομένου ότι οι ψυχικές καταστάσεις δεν μπορούν να κάνουν κάποιον να πράξει, πρέπει να δεχτούμε ότι το βουλητικό στοιχείο στην απαίτηση της πράξης αποδίδεται από το δίκαιο μόνο στα υποκείμενά του, αντανακλώντας την τάση μας να κάνουμε την ίδια απόδοση στους άλλους. Καθώς δεν μπορούμε να μπούμε στο μυαλό κάποιου, υποθέτουμε ότι αυτό που κάνει ή αυτό που δεν κάνει είναι από επιλογή.

Μέχρι εδώ, όλα καλά. Έχω υποθέσει κι εγώ το ίδιο σε αυτό το κείμενο. Ο Hage, ωστόσο, προχωρά ένα βήμα παραπέρα, υποστηρίζοντας ότι, υπό το πρίσμα των παραπάνω, τίποτα δεν μας εμποδίζει να επεκτείνουμε αυτή την απόδοση σε μη ανθρώπινους δρώντες.

> Επειδή η απόδοση συνεπάγεται έναν νου, εμπρόθετη δράση και ευθύνη μπορούν θεωρητικά να αποδοθούν σε οτιδήποτε και με οποιοδήποτε σκεπτικό. Είναι δυνατόν να θεωρούμε ορισμένα συμβάντα ως πράξεις ζώων ή θεών, ή ως πράξεις οργανισμών, και μπορούμε να κρίνουμε τα ζώα, τους θεούς και τους οργανισμούς υπεύθυνους και υπόλογους γι' αυτές τις «πράξεις». Τούτο, ωστόσο, μόνο από μια ιστορική προοπτική, κατ' αναλογία με την απόδοση εμπρόθετης δράσης στους ανθρώπους. Από οντολογική άποψη, δεν έχει διαφορά το αν αποδίδουμε εμπρόθετη δράση σε ανθρώπους ή σε άλλους δρώντες.[21]

Καθώς η απόδοση ψυχικών καταστάσεων είναι ψυχολογικό ζήτημα, *θα πρέπει* να υπάρχει μια επαρκώς διαδεδομένη ψυχολογική διάθεση αντί για τη θεσμική δυνατότητα επέκτασης του πεδίου εφαρμογής του νόμου. Αυτό ισχύει αναμφίβολα σε σχέση με τις συμπεριφορές μας προς τους άλλους ανθρώπους. Μπορεί ακόμα και να είναι επιστημονικά επαληθεύσιμο –θα μπορούσαμε να κάνουμε μια δημοσκόπηση που να επιβεβαιώνει ότι οι περισσότεροι από εμάς είμαστε της ίδιας γνώμης– αλλά το θέμα είναι ότι δεν χρειάζεται επαλήθευση. Είναι αυτό που στηρίζει τις καθημερινές μας αλληλεπιδράσεις, είτε άμεσες είτε διαμεσολαβούμενες από θεσμούς. Μπορεί να ειπωθεί το ίδιο για τη στάση μας απέναντι σε μη ανθρώπινους δρώντες; Βεβαίως, τείνουμε να εξανθρωπίζουμε τα ζώα, για παράδειγμα, και να λέμε γι' αυτά ότι ενεργούν με τρόπους παρόμοιους με τους δικούς μας, αλλά πάντα αποφεύγουμε να τους αποδίδουμε το πλήρες σύνολο των ικανοτήτων που θεωρούμε ότι έχουμε εμείς. Το ίδιο αντανακλάται και στις θεσμικές μας πρακτικές: οι περιπτώσεις στις οποίες μη ανθρώπινοι δρώντες θεωρήθηκαν ανθρώπινα υποκείμενα είναι ελάχιστες, σπάνιες και πάντα βραχύβιες.[22]

---

20. Jaap Hage, «Theoretical foundations for the responsibility of autonomous agents», *Artificial Intelligence and Law*, 25, 2017, σ. 255-271.

21. Hage, σημ. 21, σ. 261.

22. Η αναγνώριση δικαιωμάτων σε μη ανθρώπινους δρώντες είναι διαφορετικό ζήτημα, καθώς δεν κάνει τέτοιους δρώντες υποκείμενα του δικαίου αλλά μάλλον προστατευόμενούς του.

Αν το πρώτο εμπόδιο για να αποδώσουμε βουλητικά στοιχεία στις μηχανές είναι ότι *στην πράξη* δεν το κάνουμε, το δεύτερο είναι ότι δεν έχουμε ούτε καν πρόχειρες ενδείξεις ότι θα ήταν ενδεδειγμένο να το κάνουμε. Αν μη τι άλλο, έχουμε λόγους να πιστεύουμε ακριβώς το αντίθετο.

Εξ όσων γνωρίζουμε, η ενεργοποίηση ενός τεχνητού νευρώνα και η επικοινωνία του με άλλους νευρώνες είναι αυτόματες αντιδράσεις. Μπορεί να είναι σχεδιασμένες και απρόβλεπτες, αλλά αυτό απέχει πολύ από το να θεωρήσουμε ότι υπόκεινται στη θέληση της μηχανής, ότι παρακινούνται από μια ψυχική κατάσταση που διαφέρει από τη δραστηριότητά τους. Η εκπαίδευσή τους μπορεί να επιτρέπει διάφορες σειρές ενεργειών, αλλά οι επιλογές τους εξακολουθούν να καθορίζονται από τα δεδομένα. Ακόμα και τα λάθη που κάνουν, και γνωρίζουμε ότι κάνουν λάθη κατά καιρούς, ισοδυναμούν με δυσλειτουργίες στην τοποθέτηση των κομματιών ενός τεράστιου παζλ πληροφοριών, όχι με προτίμηση μιας επιλογής έναντι μιας άλλης, όπως πιστεύουμε ότι κάνουμε εμείς. Κάποιος θα μπορούσε να αντιτείνει ότι ούτε εμείς μπορούμε πραγματικά να θέλουμε να κάνουμε λάθος, αλλά η ένσταση θα ήταν άστοχη. Το ζήτημα εδώ δεν είναι αν η βούληση υπόκειται σε δεοντολογικούς περιορισμούς, αλλά αν οι μηχανές έχουν τη βασική ικανότητα να αναπτύσσουν μια αποστασιοποιημένη στάση σε σχέση με τις «πράξεις» τους.

Οι μηχανές δεν έχουν ούτε την ικανότητα να απεικονίζουν το περιβάλλον τους, όπως αναμένει το ποινικό δίκαιο από τα υποκείμενά του. Η αντίληψή τους γι' αυτό είναι αποσπασματική – πρέπει να αναλύουν καθετί σε μικροσκοπικά συστατικά μέρη, και όταν το ανασυνθέτουν ως δεδομένα εξόδου, δεν έχουν μια ολιστική γνώση του, όπως θα είχαμε εμείς με την κοινωνική σημασία του πράγματος.

Φανταστείτε ότι ένας αναλυτής για έναν χορηγό ενυπόθηκων δανείων, ας τον πούμε Άνταμ, έχει αναλάβει να προβλέψει ποιοι από τους υπάρχοντες δανειολήπτες είναι πιθανό να καθυστερήσουν τις πληρωμές τους ή να αθετήσουν τις υποχρεώσεις τους. Ο Άνταμ εισάγει σε έναν υπολογιστή όλες τις πληροφορίες που θεωρούνται σχετικές, όπως ηλικία, διεύθυνση, επαγγελματικό ιστορικό και επαγγελματική κατάσταση, πιστωτικό ιστορικό, κ.ο.κ., καθώς και γενικά ιστορικά αρχεία αθέτησης υποχρεώσεων. Ας υποθέσουμε ότι, λόγω κάποιας τεχνικής δυσλειτουργίας, οι εισφορές κοινωνικής ασφάλισης της δανειολήπτριας Μπέτι δεν καταχωρούνται και, ως εκ τούτου, εμφανίζεται άνεργη για ένα έτος. Λαμβάνοντας αυτό υπόψη, ο υπολογιστής καταλήγει στο συμπέρασμα ότι η Μπέτι είναι πιθανό να αθετήσει τις υποχρεώσεις της εντός του επόμενου τριμήνου. Στη συνέχεια, ο Adam διαβάζει τα αποτελέσματα και τα στέλνει με email στον προϊστάμενό του.

Κατ' αρχάς, ο υπολογιστής δεν γνωρίζει ότι κάνει μια αναπαράσταση, ότι οι υπολογισμοί του αντανακλούν μια κατάσταση πραγμάτων έξω από αυτόν, πόσο μάλλον ότι υπάρχει ένα πραγ-

ματικό πρόσωπο με το όνομα Μπέτι στο οποίο αναφέρονται οι εν λόγω υπολογισμοί και του οποίου το μέλλον θα επηρεαστεί από αυτούς. Ούτε γνωρίζει ότι οι υπολογισμοί κοινοποιούνται σε κανέναν. Τίποτα δεν υπάρχει έξω από τα ΤΝΔ του. Ο υπολογιστής αντιλαμβάνεται το περιβάλλον του μόνο όταν το εσωτερικεύει ως δεδομένα. Ο Άνταμ, από την άλλη πλευρά, είναι σε θέση, ή έτσι υποθέτει το ποινικό δίκαιο, να τοποθετήσει τη δήλωσή του σε ένα πλαίσιο που υπάρχει έξω από αυτόν, να της προσδώσει νόημα. Για τον ίδιο λόγο, ο υπολογιστής δεν είναι σε θέση να παραπλανήσει κανέναν, καθώς, σε ό,τι τον αφορά, δεν υπάρχει κανένας για να παραπλανήσει. Γι' άλλη μια φορά, δεν ισχύει το ίδιο για τον Άνταμ,. Η αναπαράσταση της μηχανής δεν μπορεί επίσης να είναι ποτέ ψευδής. Ό,τι της εισάγεται, ακόμα και τις λανθασμένες πληροφορίες σχετικά με την κοινωνική ασφάλιση της Μπέτι, δεν μπορεί παρά να τις θεωρήσει αληθές, παρότι μπορεί να αποδώσει διαφορετική αξία σε διαφορετικά σύνολα πληροφοριών, καθώς δεν έχει την ικανότητα να τις αντιπαραβάλει, μάλιστα ούτε καν να σκεφτεί να τις αντιπαραβάλει, με οποιαδήποτε εναλλακτική.

Αν ισχύουν όλα αυτά, τότε δεν υπάρχουν ποτέ συνθήκες υπό τις οποίες δύνανται οι υπολογιστές να πληρούν την actus reus στην έποψή της ως ψυχικής κατάστασης.

Υπάρχουν δύο πιθανά αποτελέσματα του ελλείμματος φαντασίας της μηχανής.

Πρώτον, θα μπορούσε να θεωρηθεί ότι θέτει σε κίνδυνο την ικανότητα της μηχανής να της αποδοθεί ποινική ευθύνη. Κάποιος αποκλείεται από την ποινική ευθύνη όταν δεν μπορεί, ή θεωρείται ότι δεν είναι ικανός, όπως κατεξοχήν τα παιδιά στις περισσότερες έννομες τάξεις, να διακρίνει μεταξύ σωστού και λάθους. Η ίδια αρχή κανονικά επεκτείνεται, τουλάχιστον φιλοσοφικά, αν όχι θεσμικά, σε όσους δεν έχουν την ικανότητα να κατανοήσουν τις συνέπειες των πράξεών τους και, ακόμα περισσότερο, σε όσους δεν είναι σε θέση να τοποθετήσουν *ολωσδιόλου* τις πράξεις τους σε ένα πλαίσιο. Αλλά ακόμα κι αν δεν δεχθούμε αυτή την επέκταση, επειδή δεν έχει επαρκή θεσμική υποστήριξη, η πλήρης αδυναμία να κατανοήσουν το πραγματικό πλαίσιο θα τους παρείχε τουλάχιστον τη δικαιολογία της παραφροσύνης. Αν ακολουθήσουμε αυτή την κατεύθυνση, δεχόμαστε ότι η δραστηριότητα των μηχανών μπορεί να πληροί την actus reus των αδικημάτων, αλλά, δεδομένης της φύσης τους, είναι πρακτικά αδύνατο να θεωρηθούν οι μηχανές ποινικά υπεύθυνες.

Δεύτερον, θα υποστήριζα, ενδεχομένως κάπως πιο φιλόδοξα, ότι το έλλειμμα φαντασίας της μηχανής ακυρώνει τον ίδιο τον χαρακτήρα της δραστηριότητάς της ως αξιόποινης.

Όποια στάση κι αν υιοθετεί κανείς ως προς την κανονιστικότητα του δικαίου, είτε πιστεύει ότι είναι ανεξάρτητη από τις κοινωνικές συμπεριφορές είτε θεωρεί ότι είναι θέμα ευρείας κοινω-

νικής αποδοχής, πρέπει να δεχτεί ότι η καταληπτότητα είναι, αν μη τι άλλο, προϋπόθεση της ύπαρξής του. Το δίκαιο πρέπει να δέχεται τα υποκείμενά του και τον κόσμο όπως είναι. Για να είναι ο νόμος κατανοητός, πρέπει προφανώς να εκφράζεται σε μια κατανοητή γλώσσα. Δεύτερον, το νόημα των νομικών εννοιών μπορεί να εξειδικεύεται για κάθε σύστημα και να καθορίζεται από τη νομική πρακτική, αλλά πρέπει μολαταύτα να βασίζεται στη και να συνάδει με τη λαϊκή σημασία των λέξεων. Τρίτο και σημαντικότερο για τους σκοπούς μας, το δίκαιο πρέπει να μπορεί να συλλαμβάνει τον κόσμο με νοητό τρόπο.

Για να επαναλάβουμε όσα είπαμε λίγο παραπάνω, αυτό προϋποθέτει ικανότητα φαντασίας εκ μέρους των υποκειμένων του δικαίου. Σημειωτέον ότι το κρίσιμο είναι η γενική ικανότητα, όχι η δυνατότητα της σωστής άσκησής της. Επομένως, ακόμα και όσοι έχουν προσωρινά ή μόνιμα μειωμένη ικανότητα να προβλέπουν τις συνέπειες των πράξεών τους ή να αποδίδουν σε αυτές ένα γενικά αποδεκτό νόημα, εξακολουθούν να κατέχουν τη γενική ικανότητα.

Επιτρέψτε μου να επιστρέψω στον Hage. Θυμηθείτε ότι υποστηρίζει πως, εφόσον είναι θεσμικά δυνατόν να αποδίδονται στις μηχανές προθετικές καταστάσεις, το πραγματικό ερώτημα είναι αν είναι επιθυμητό αυτό. Ελλείψει οποιασδήποτε απόδειξης ότι οι μηχανές πράγματι παίρνουν από τις ενέργειές τους την απόσταση που τα συνειδητά υποκείμενα θεωρούν ότι παίρνουν, η απόδοση αυτή θα ήταν ένα νομικό πλάσμα, θα αντιμετωπίζαμε τις μηχανές *σαν να* διέθεταν τις απαιτούμενες ικανότητες. Γιατί, θα μπορούσε να ρωτήσει κανείς, δεν μπορούμε να δεχτούμε ότι, εφόσον το δίκαιο είναι κατανοητό σε επαρκή αριθμό υποκειμένων, διατηρεί την κανονιστικότητά του όχι μόνο έναντι αυτών αλλά και έναντι όσων δεν πληρούν τις προϋποθέσεις της καταληπτότητας;[23] Διότι κάτι τέτοιο θα ήταν αντιφατικό. Όταν ο νόμος ισχυρίζεται ότι δεσμεύει τον Α, ισχυρίζεται ότι το κάνει σε έναν πιθανό κόσμο, στον οποίο όλα τα υποκείμενά του φέρουν τα ίδια ακριβώς χαρακτηριστικά με τον Α. Φανταστείτε τώρα έναν πιθανό κόσμο, στον οποίο όλα τα υποκείμενα φέρουν τα χαρακτηριστικά μιας μηχανής ΤΝ, συμπεριλαμβανομένου του ελλείμματος φαντασίας. Σε αυτόν τον κόσμο, το δίκαιο θα ήταν στην πραγματικότητα εντελώς ακατανόητο, επομένως η κανονιστικότητά του θα υπονομευόταν θανάσιμα. Ένα νομικό σύστημα που θεωρεί τις μηχανές ΤΝ ενδεδειγμένα υποκείμενά του θα ήταν, ταυτόχρονα, ανίκανο να έχει τις μηχανές ΤΝ ως ενδεδειγμένα υποκείμενά του. Έπεται ότι το δίκαιο δεν μπορεί να δεσμεύει τις μηχανές ΤΝ ως υποκείμενα. Κατά συνέπεια, οι «πράξεις» τους δεν μπορούν ποτέ να πληρούν τις απαιτήσεις της actus reus.

23. Εδώ εγείρεται προφανώς το σχετικό με τη δικαιοσύνη ερώτημα αν μια έννομη τάξη με αποστολή να καθοδηγεί τη δράση ενδέχεται να μην καταφέρει να καθοδηγήσει τις πράξεις των υποκειμένων της – ωστόσο δεν μας αφορά άμεσα στο παρόν πλαίσιο.

Το κενό ευθύνης

Ας θυμηθούμε γιατί πολλοί ανησυχούν για το κενό της τεχνο-ευθύνης. Οι μηχανές ΤΝ μπορούν να προκαλέσουν βλάβη. Όταν το κάνουν, οι ίδιες δεν μπορούν να θεωρηθούν ποινικά υπεύθυνες, επειδή δεν έχουν την ικανότητα να σχηματίσουν mens rea. Ούτε οι προγραμματιστές είναι ποινικά υπεύθυνοι, στο βαθμό που έχουν εγκαταλείψει τον έλεγχο των μηχανών. Αυτό μας αφήνει απογοητευμένους, διότι δεν υπάρχει κανείς στον οποίο να μπορούμε να κατευθύνουμε τις αντιδράσεις μας για κάτι που μας φαίνεται άδικο και όχι απλό ατύχημα.

Ποιος είναι ο αντίκτυπος στο κενό ευθύνης, αν όντως ισχύει ότι οι μηχανές δεν μπορούν να ενεργήσουν με τρόπο που να άπτεται του ποινικού δικαίου; Εκ πρώτης όψεως, το κενό μοιάζει να διευρύνεται. Η εντύπωση του άδικου και το αίσθημα ότι η βλάβη είναι τυχαία δεν εξαφανίζονται, και οι αντιδράσεις μας που τείνουν στην απόδοση ευθυνών επιμένουν. Το γεγονός ότι δεν υπάρχει καθόλου πράξη ή δράστης κάνει την απογοήτευση ακόμα πιο έντονη, διότι μας μένει η αίσθηση ότι η τροποποίηση της θεσμικής μας τάξης, ώστε να διορθωθούν τέτοιου είδους αδικίες, είναι ακόμα πιο δύσκολη.

Θα έλεγα ότι πρόκειται για εσφαλμένη εντύπωση. Αυτό που φαινομενικά γεννά την ευθύνη είναι κατ' αρχάς η ενδιάμεση «πράξη» της μηχανής, η οποία απαλλάσσει τους προγραμματιστές ή τους ιδιοκτήτες τουλάχιστον για τη βλάβη που προκλήθηκε στο τέλος της διαδικασίας (κάτι που προφανώς δεν αποκλείει την ευθύνη για άλλα ποινικά αδικήματα που διαπράχθηκαν ανεξάρτητα σε προγενέστερο στάδιο).Υποστήριξα, ωστόσο, ότι οι μηχανές δεν ενεργούν με τρόπο που άπτεται του ποινικού δικαίου, καθιστώντας συνεπώς αδύνατη την απόδοση αξιόποινων πράξεων σε αυτές ή την εξατομίκευσή τους ως δράστες. Αν ισχύει αυτό, τότε καταργείται η ενδιάμεση ζώνη μεταξύ των πράξεων του προγραμματιστή και της τελικής βλάβης.

Ανοίγονται δύο επιλογές. Η πρώτη είναι να δεχτούμε ότι οι μηχανές ενεργούν μεν, αλλά πάντα, χωρίς εξαίρεση, ως αυτόματα ή χωρίς την ικανότητα να αξιολογούν τη φύση και το νόημα των πράξεών τους. Ο νόμος θα πρέπει τότε να τις αντιμετωπίζει ως ανυπαίτιους δρώντες. Αν πληρούνται οι σχετικές προϋποθέσεις, τότε η ευθύνη θα επιστρέφει στους προγραμματιστές ή στους χρήστες, οι οποίοι θα θεωρηθεί ότι διέπραξαν το αδίκημα μέσω ενός ανυπαίτιου δρώντα.[24] Είναι ένα ενδεχόμενο αυτό, αλλά ταυτόχρονα θα είχαμε τότε τη μάλλον δυσάρεστη κατάσταση να υπάρχει μια ολόκληρη κατηγορία υποκειμένων του ποινικού δικαίου η οποία, ωστόσο, δεν δύναται *ποτέ* να πληροί τις προϋποθέσεις της ποινικής ευθύνης.

24. Αυτό είναι ένα από τα μοντέλα ευθύνης που προτείνει ο Hallevy. Η διαφορά είναι πως, επειδή υποθέτει ότι οι προηγμένες μηχανές όντως ενεργούν με τρόπο που άπτεται του ποινικού δικαίου, επιφυλάσσει αυτή την επιλογή για λιγότερο προηγμένους υπολογιστές.

Η νομική ευθύνη θα επιστρέφει και πάλι στους προγραμματιστές ή στους χρήστες, αν δεχτούμε ότι η δραστηριότητα της μηχανής δεν μπορεί να θεωρηθεί επ' ουδενί actus reus, μόνο που αυτή τη φορά δεν υπάρχει τίποτα που να τους διαχωρίζει από τη βλάβη ή τον κίνδυνο βλάβης που προκαλείται.

Οι προϋποθέσεις θα ποικίλλουν, φυσικά. Το γεγονός ότι οι μηχανές παραμένουν αδιαφανή «μαύρα κουτιά» εξακολουθεί να έχει σημασία, καθορίζοντας κρίσιμες όψεις της ευθύνης, όπως η απόσταση της βλάβης από τις πράξεις του προγραμματιστή ή του χρήστη, η σοβαρότητα του κινδύνου, κ.ο.κ. Όσες δυσκολίες όμως κι αν υπάρχουν στη δίκαιη κατανομή της ποινικής ευθύνης, κενό τεχνο-ευθύνης δεν υπάρχει.