«L'etica dei robot killer. Sviluppi, caratteristiche e conseguenze della guerra artificiale»
Scuola Critica del Digitale – Forum Disuguaglianze e Diversità
Mercoledì 25 November 2020

@guidonld
@LawScottish

Robot killer ed 'ethicswashing'

Prof. Avv. Guido Noto La Diega
Associato di Proprietà Intellettuale e Privacy

Due trend interconnessi

- Investimenti in armi letali automatiche / robot killer (Russia, Israele)
  - '*weapon system that, once activated, can select and engage targets without further intervention by a human operator*' (US Department of Defence Directive 3000.09)
- Ethics by design/value-sensitive design/ethical AI – civile e militare (Umbrello, Arkin, BSI, etc.)
  - Emotionally intelligent AI (D'Mello, Shibata, Schuller, etc.)
  - Artificial consciousness (Baars, Aleksandr, Warren, etc.)

The Case for Ethical Autonomy in Unmanned Systems

Ronald C. Arkin

Moral Decision Making in Autonomous Systems: Enforcement, Moral Emotions, Dignity, Trust, and Deception

/ Ronald Craig Arkin, *Fellow IEEE*, Patrick Ulam, and Alan R. Wagner

Ethical Robots in Warfare

RONALD C. ARKIN

Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture PART I: Motivation and Philosophy

Ronald C. Arkin
Mobile Robot Laboratory
Georgia Institute of Technology

‘[A]n unmanned system will be able to be perfectly ethical in the battlefield (…) **they can perform more ethically than human soldiers are capable of**’ (Arkin 2007, 4)

BS 8611:2016

BSI Standards Publication

Robots and robotic devices

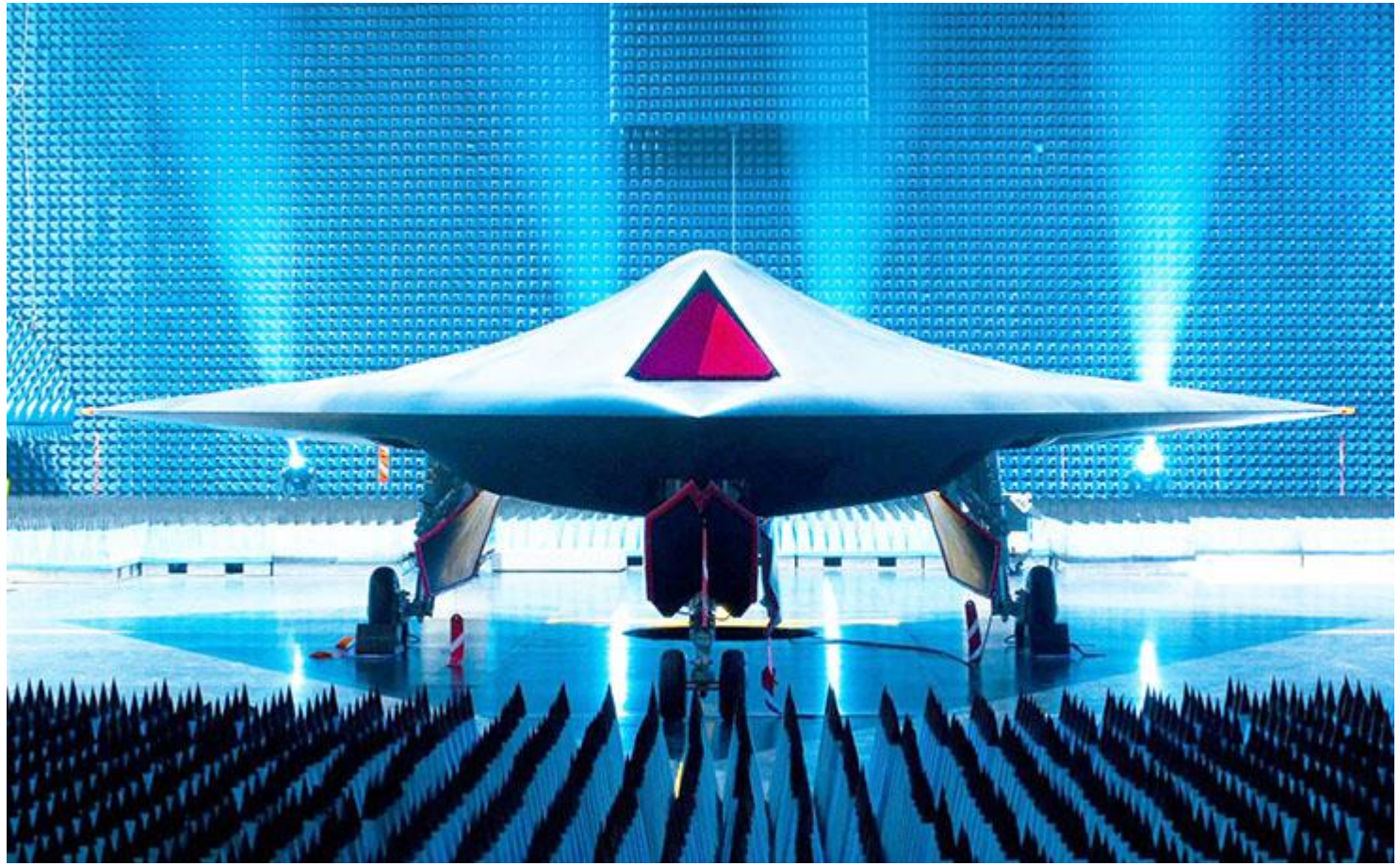
Guide to the ethical design and application of robots and robotic systems

La terza rivoluzione bellica

Polvere da sparo -> armi nucleari -> **armi letali autonome (LAWs)**

I sistemi in utilizzo sono ad autonomia **ristretta**

Iron Dome system israeliano Lancia missile automaticamente, di fatto senza lasciare tempo per **intervento umano**

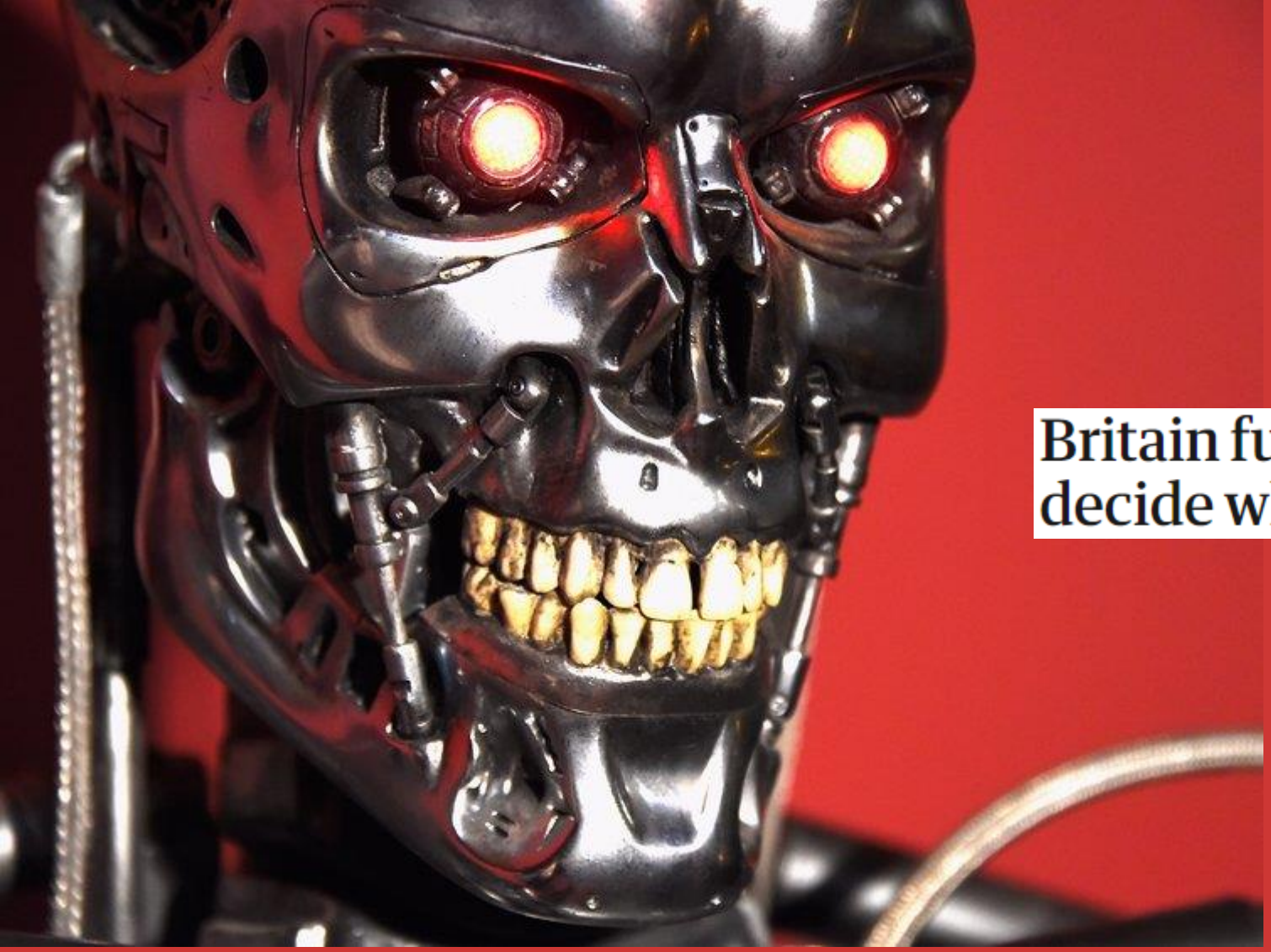


Una guerra “pulita”

- I robot killer sono tanto in grado di comportarsi moralmente e rispettare il *jus in bello*

- **Niente piu’ spargimenti di sangue**

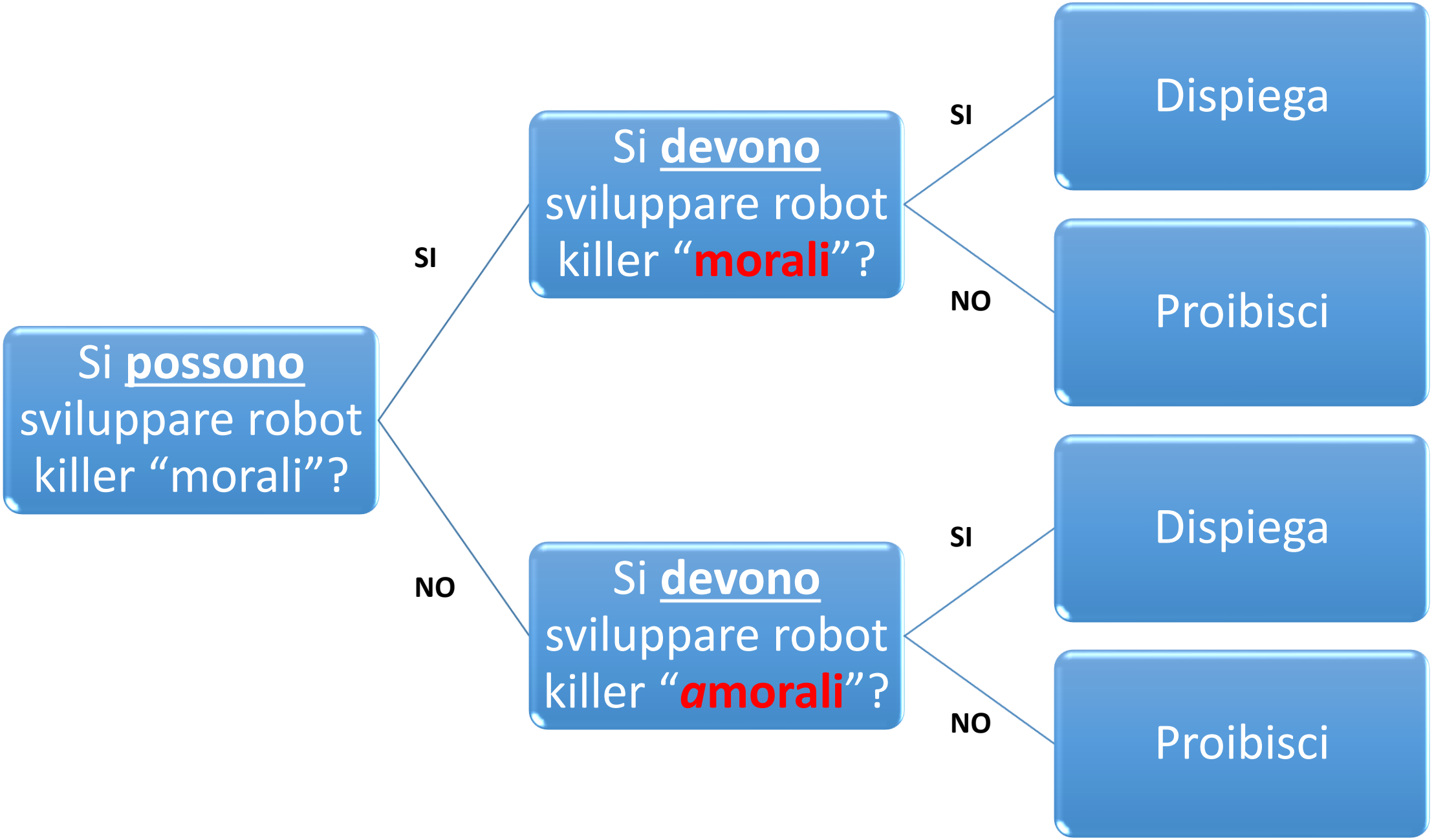
- **I nostri soldati saranno al sicuro, la nostra nazione sara’ piu’ forte**



Politicamente corretti?

Britain funds research into drones that decide who they kill, says report

'We believe a preemptive ban is premature' (UK Ministry of Defence, 29 March 2019)



Scrivere l'etica nel codice dei robot killer

Robot killer '*capable of performing **more ethically** on the battlefield **than (…) human soldiers***' (Arkin 2010)

Possono agire conformemente ai principi di **distinzione, proporzionalita', necessita'** (Arkin 2008)

(i) Le decisioni dei robot non sono influenzate dalle **emozioni**

(ii) I loro sensori sono migliori degli umani, il che migliora la performance sul campo di battaglia

Il “governatore etico” di Arkin

- **Collo di bottiglia** che consente solo azioni eticamente accettabili

- Vincoli basati su **rappresentazioni dei principi di diritto** umanitario e regole d’ingaggio

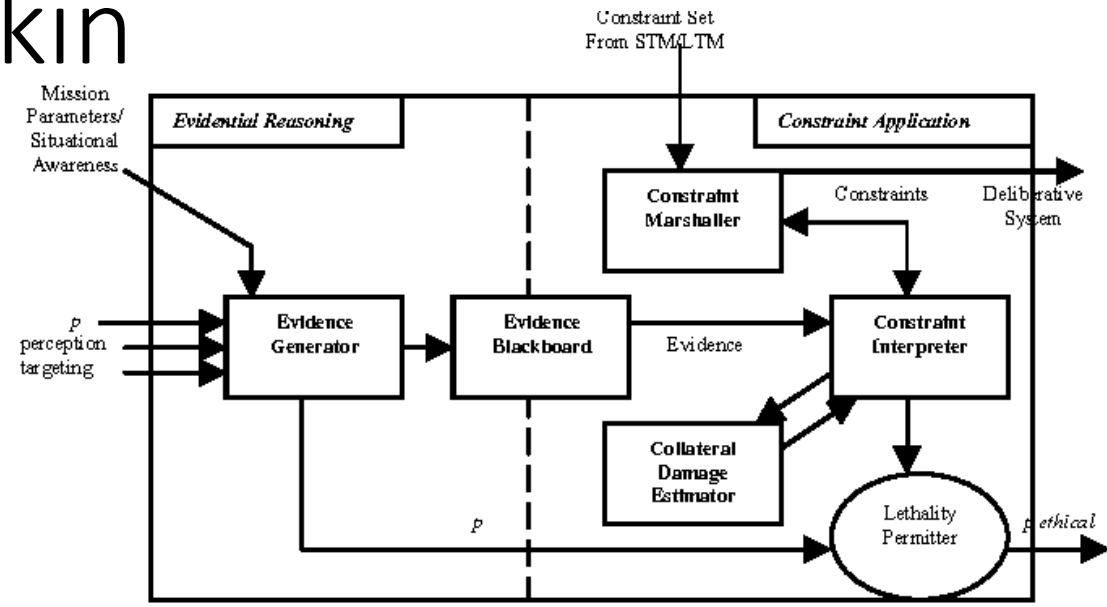


Figure 3. Architecture and data flow overview of the ethical governor

- Nell’**ambiente di simulazione** ‘MissionLab’ scenari di prova: decisioni circa l’uso della forza limiterebbero i danni collaterali

Contro Arkin

- **Consulente** etico, non “governatore” (i soldati possono scavalcarlo) (Matthias 2011)

- Presuppone una dubbia concezione dell’agente morale come qualcuno che **segue ciecamente le regole** (Johnson and Axinn 2013)

- Puo’ essere programmato anche per prendere **decisioni immorali** (Vanderelst and Winfield 2016)

- **I principi** di diritto umanitario e le regole d’ingaggio sono **vaghi e contraddittori** (Matthias 2011, Noto La Diega 2019)

Distinzione

- Il diritto internazionale umanitario protegge la popolazione civile e vieta attacchi a civili e a beni civili
- Parti in conflitto devono rivolgere i loro attacchi militari esclusivamente contro obiettivi militari e devono di conseguenza sempre distinguere tra **civili e combattenti** e tra **beni civili e obiettivi militari** (I Protocollo alle Convenzioni di Ginevra, artt 48, 51(2), 52(2))
- Non si possono attaccare i combattenti se '***hors de combat'*** (Customary Int'l Humanitarian Law, Rule 47)
- 'Il principio della distinzione porta a una limitazione dei metodi e dei mezzi di combattimento: tutte le armi o strategie che non sono impiegate in modo mirato contro un obiettivo militare sono vietate' (DFAE 2018)

Distinzione

- **Una madre spaventata** che corre dietro i suoi bambini e grida loro di smettere di giocare con le pistole giocattolo potrebbe essere interpretata come qualcuno che corre verso due persone armate e quindi un **obiettivo legittimo** (HRW 2012)
- Presuppone la **compresione delle intenzioni** alla base delle decisioni umane

Proporzionalità

- Attacchi sproporzionati se “ci si può attendere che provochino incidentalmente morti e feriti fra la popolazione civile, danni ai beni di carattere civile, o una combinazione di **perdite umane e di danni**, che risulterebbero **eccessivi rispetto al vantaggio militare concreto** e diretto previsto” (I Protocollo, art 51(5)(b))
- Proporzionalità richiede “*responsible accountable human commanders, who can weigh the options based* ***on experience and situational awareness***” (Noel Sharkey 2012a, b; Suchmann 2016)

SE *n* bambini
ALLORA uccidi 1 terrorista
n = ?

Necessita’

- **Forza militare** andrebbe usata solo nella misura necessaria per vincere la guerra (Schmitt 2010)

- Intrinsecamente **umano**: valutazione e’ giudizio di valore in cui il comandante deve **bilanciare gli imperativi della vittoria e il requisito di umanita’** (Kastan 2013)
- Non appena LAWs “*are widely introduced, it becomes a matter of military necessity to use them, as they* ***could prove far superior*** *to any other type of weapon*” (Krishnan 2009, 91)

Compresione umana e teatri di guerra

- Applicare distinzione, proporzionalita' e necessita' in situazioni reali ≠ tradurli in **codice binario** sulla base di **ipotesi di laboratorio**

- Comprendere il mondo e' un processo olistico e discrezionale – non si presta a **riduzioni algoritmiche**

- Nei teatri di guerra, l'interpretazioni richiede **empatia e compassione**, quintessenzialmente umane

Clausola Martens

- Possibile obiezione: non esiste un trattato internazionale che vieti espressamente i LAWS

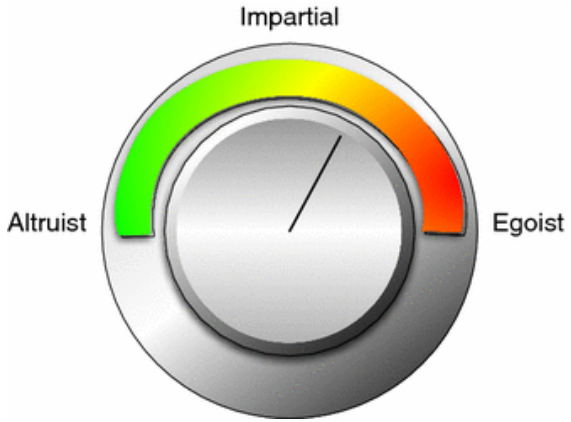
- La clausula Martens proibisce qualsiasi arma che sia contraria alle “esigenze della coscienza pubblica” (preamboli alle Convenzioni dell’Aja del 1899 e del 1907)

Irresponsabilita’

- Anche gli operatori umani faticano ad applicare distinzione, proporzionalita’ e necessita’, ma **la violazione e’ perseguita** dinanzi ai tribunali militari
- Robot killer **non possono essere chiamati a rispondere delle violazioni**
- Non e’ chiara in che circostanze la responsabilita’ possa cadere sul **comandante** che ne ha ordinato l’uso, il **programmatore** o il **costruttore**
- Considerare queste machine come ‘morali’ rinforza la tendenza degli uomini a **non assumersi le loro responsabilita**’ (Johnson 2006; Sharkey 2017)

Come si sviluppano robot ‘morali’?

1. **Li si programma** onde agiscano moralmente
   1. Ethical governor (Arkin 2007, 2009)
   2. Hole-avoiding robots (Winfield et al. 2014)
   3. MedEthEx (Anderson et al. 2006)



Contissa, Lagioia, and Sartor 2017

2. **Li si allena** affinche’ sviluppino moralita’
   - Pharmacy World: allineamento valoriale ottenuto mediante la lettura di storie e il reverse engineering dei relative valori (Riedl and Harrison 2015)

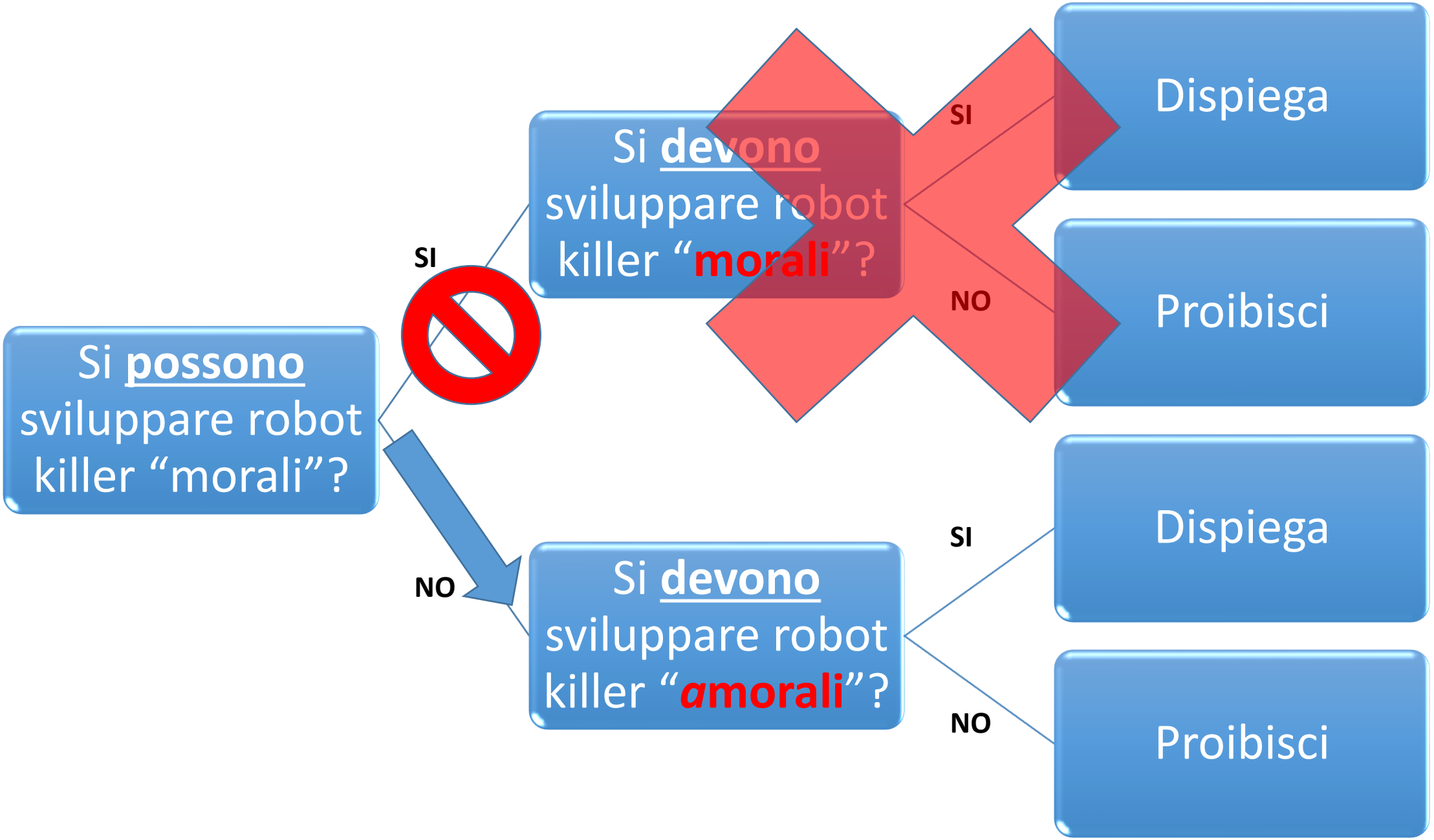
Difetti

- Approccio ‘top-down’ (programmarli, Arkin): 1) sperimentazione in **contesti simulati** e applicazioni limitate; 2) chi decide **cosa sia ‘morale’**? 3) Se ci accordiamo che IHL e’ sintesi di principi morali cardine, e’ possibile **tradurli in codice binario**?

- Approccio dal basso (training, Riedl and Harrison): imparano osservando e interagendo con gli esseri umani. 1) Si puo’ replicare su **larga scala**? 2) Vogliamo **interagire coi robot killer** per permetter loro di imparare i nostri valori?

Quali alternative?

- Non esistono robot morali stricto sensu
- "*We can't say for certain that **future** machines will lack [conosciousness, intentionality and free will]*" (Moor 2006, 20)
- Opzione 1: **investire nello sviluppo di competenze morali**
  - Per minimizzare i rischi delle macchine autonome (Wallach 2010)
  - Per migliorare la nostra comprensione dell'etica (Moor 2006)
- Opzione 2: **non usare i robot i contesti che richiedono competenze morali**
  - I robot non dovrebbero essere autorizzati a decidere se uccidere perche' non hanno "*human judgement, common sense, appreciation of the larger picture, understanding of intentions behind people's actions*" (Heyns 2013).



Si **devono** sviluppare robot killer "*a*morali"?

1. Sono necessari per ridurre il numero eccessivo di morti fra i soldati?

2. Sono piu' efficaci dei soldati?

Sono necessari per ridurre il numero eccessivo di morti fra i soldati?

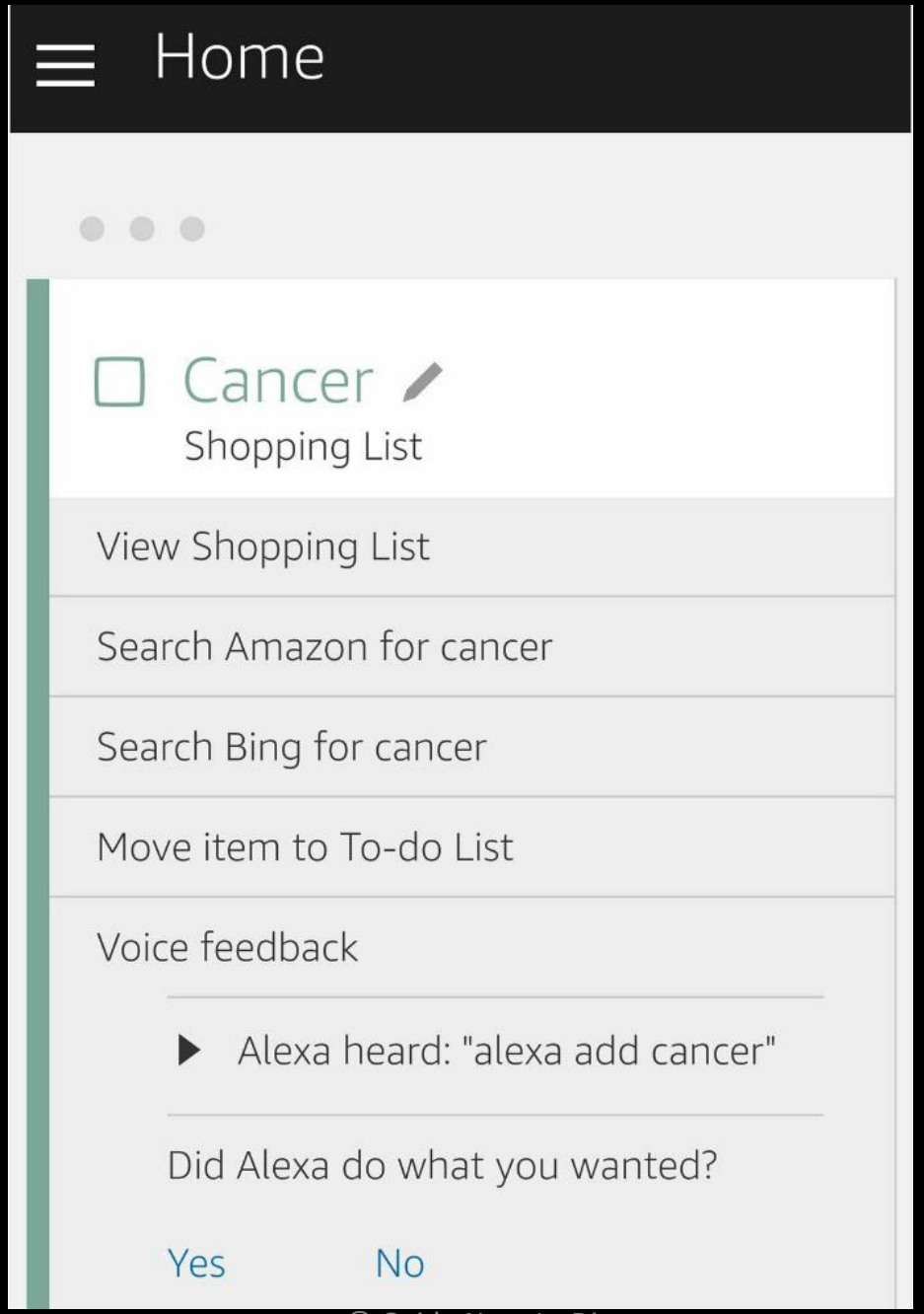
- Nel 2019, 1 solo militare italiano e' morto in missione (per un malore)
- In Regno Unito, su 61, 16% cancro, 16% incidenti stradali, 36% altri incidenti (16/22 suicidi)
- **Solo 1** e' morto a seguito di un attacco nemico
- "*UK Regular Armed Forces were at a statistically significant **lower risk** of dying compared to the UK **general population***" (ONS 2020)

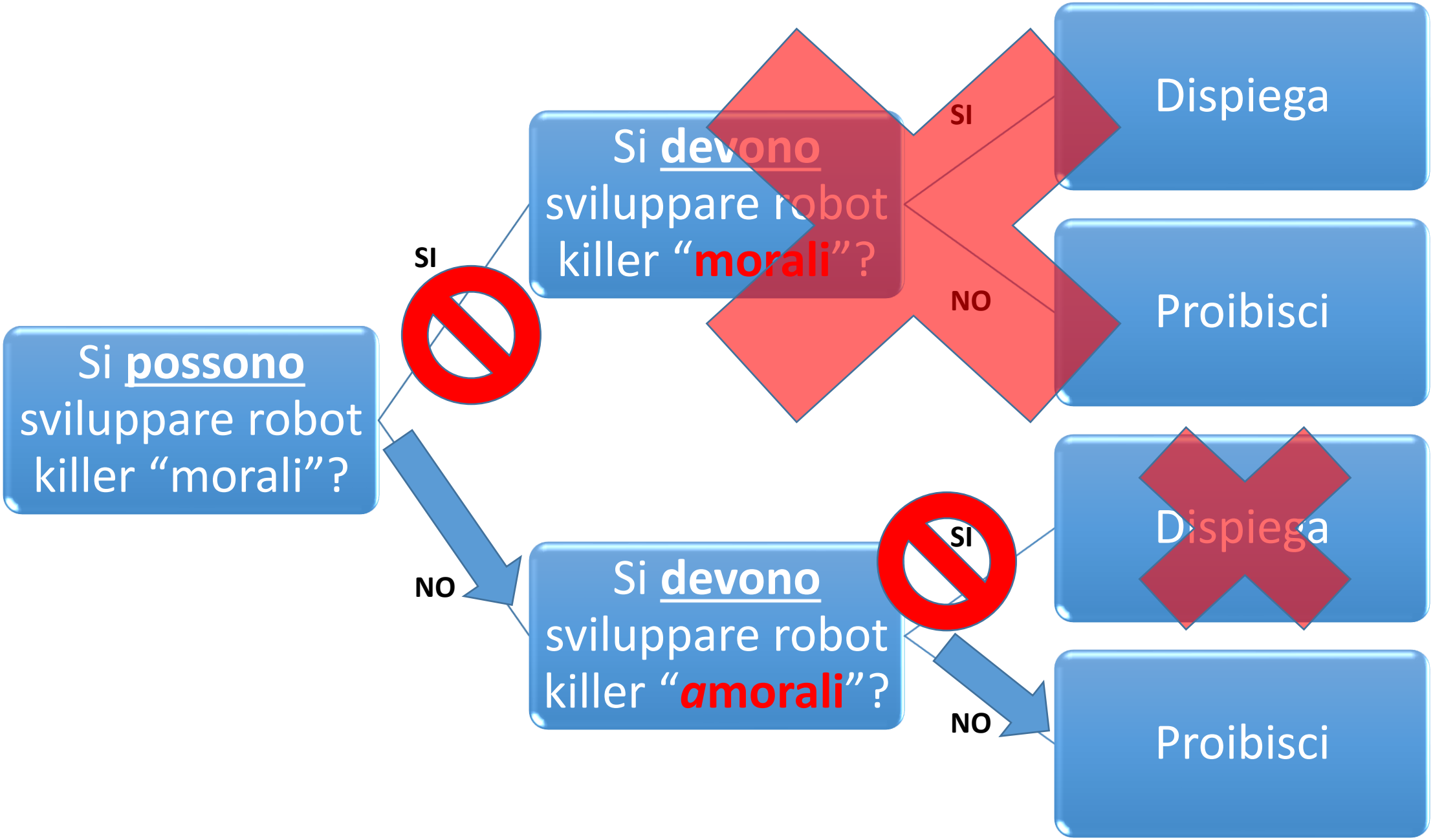
Sono necessari per ridurre il numero eccessivo di morti fra i soldati?

- **Piu' incentive a cominciare una guerra** (Noel Sharkey 2011)
- Forte **attaccamento emotivo** fra soldato e robot < efficace (Carpenter 2013)
- **Autonomia** aumenta errori fatali

Sono piu’ efficaci dei soldati?

1. Vulnerabilita’
2. Interazioni accidentali
3. Realta’ tecnologica Vs fantascienza

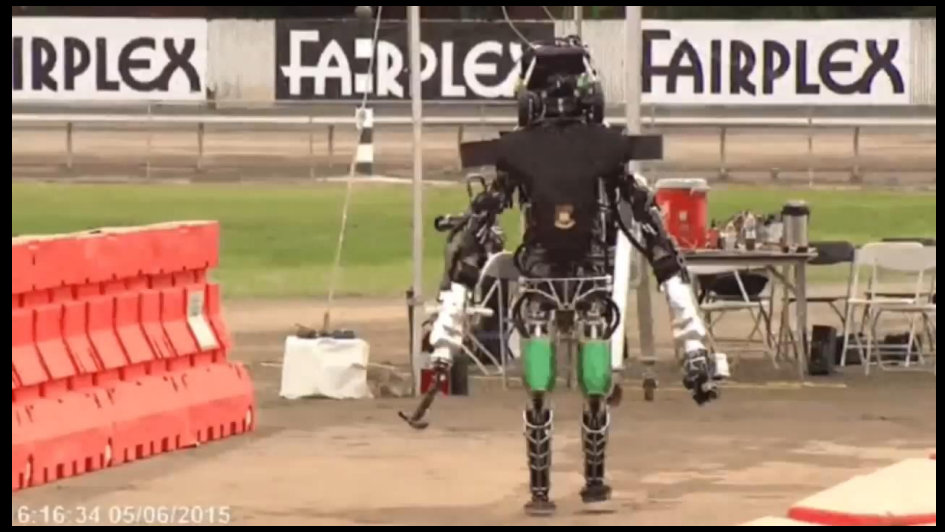




Una conclusione non-binaria

- Continuiamo a investire nella traduzione dell'etica in codice binario in **settori a basso rischio** (e.g. Roomba)
- Investiamo in **Ricerca & Sviluppo** etici/responsabili
- Lo sviluppo di robot killer **morali non** e' **possibile**
- Lo sviluppo di killer robot **amorali** e' possibile ma **non desiderabile**: inefficaci e pericolosi
- Artificial conscience/ethics by design e' serve a **confondere le acque**, rendere l'uso dei LAWs piu' accettabile e rifuggire ogni forma di responsabilita'

Grazie!



Restiamo connessi

Guido Noto La Diega
www.guidonotoladiega.com
gn12@stir.ac.uk
@guidonld

Cenni bibliografici

- Arkin, R. C. (2007). Governing lethal behaviour: Embedding ethics in a hybrid deliberative/reactive robot architecture. Atlanta: Georgia Institute of Technology.
- Asaro, P. (2012). On banning autonomous lethal systems: Human rights, automation and the dehumanizing of lethal decision-making, special issue on new technologies and warfare. *International Review of the Red Cross, 94*(886), 687–709
- Hew, P. C. (2014). Artificial moral agents are infeasible with foreseeable technologies. *Ethics and Information Technology, 16*, 197–206.
- Heyns, C. (2017). Autonomous weapons in armed conflict and the right to a dignified life: An African perspective. *South African Journal on Human Rights, 33*(1), 46–71.
- Human Rights Watch. Losing Humanity: The Case against Killer Robots. *HRW* 29 November 2012.
- Human Rights Watch (2018). Heed the Call. A Moral and Legal Imperative to Ban Killer Robots.
- Johnson, D. G., & Miller, K. W. (2008). Un-making artificial moral agents. *Ethics and Information Technology, 10*, 123–133.
- Lin, P., Abney, K., & Bekey, G. A. (2014). *Robot ethics: the ethical and social implications of robotics*. The MIT Press.
- Matthias, A. (2011). Algorithmic moral control of war robots: Philosophical questions. *Law, Innovation and Technology*, *3*(2), 279-301.
- Guido Noto La Diega, 'The artificial conscience of lethal autonomous weapons: marketing ruse or reality?' (2018) 1 Law and the Digital Age 1-17
- Sharkey, A. (2017). Can we program or train robots to be good? *Ethics and Information Technology*. https://doi.org/10.1007/s10676-017-9425-5