Responsibility Gaps and Retributive Dispositions: Evidence from the US, Japan and Germany

Markus Kneer[1]
Markus Christen
Digital Society Initiative
University of Zurich

Draft, March 2023.

Danaher (2016) has argued that increasing robotization can lead to *retribution gaps*: Situation in which the normative fact that nobody can be justly held responsible for a harmful outcome stands in conflict with our retributivist moral dispositions. In this paper, we report a cross-cultural empirical study based on Sparrow's (2007) famous example of an autonomous weapon system committing a war crime, which was conducted with participants from the US, Japan and Germany. We find that (i) people manifest a considerable willingness to hold autonomous systems morally responsible, (ii) partially exculpate human agents when interacting with such systems, and that more generally (iii) the possibility of normative responsibility gaps is indeed at odds with people's pronounced retributivist inclinations. We discuss what these results mean for potential implications of the retribution gap and other positions in the responsibility gap literature.

Keywords: Responsibility gap, autonomous weapon systems, artificial intelligence, retribution, robotics

1.Introduction

A proper understanding of the looming threat of responsibility gaps in the use of autonomous systems has several levels: (1) The *moral-philosophical question* as to who, if anyone, can be justly held responsible for harm brought about in certain human-robot interactions. (2) The *moral-psychological question*

[1] Corresponding author: Markus.kneer@gmail.com.

about actual human dispositions to attribute responsibility in such contexts. (3) The *legal, political, and societal implications* for the use of autonomous systems and how they should be regulated. In an interesting recent paper exploring all three levels of the question, Danaher (2016) has discussed the possible divergence between people's retributivist nature and the impossibility of holding anybody justly responsible. Here we explore such "retribution gaps" in a cross-cultural empirical study with participants from the US, Japan and Germany. Evidence of this sort, we argue, is of key importance for the discussion of the possible implications of retribution gaps.

1.1 Control & Responsibility

Moral culpability standardly requires agents to have a certain measure of control over their actions and outcomes. A driver, whose wheel comes off while driving, is blameless for the ensuing damages – at least if she drives responsibly, has the car checked regularly and if the conundrum was unforeseeable. The *Control Principle* is old hat in moral philosophy. It figures, perhaps, most prominently in debates about moral luck (Williams, 1981; Nelkin, 2004) and is sometimes traced back to Kant (1998/1758), though it has certainly been tacitly assumed in ethics going back to the Ancient Greeks. What is more, the *Control Principle* is a central pillar of Western Criminal Law, which discourages the punishment of unlucky accidents ("strict liability", see e.g. Fletcher, 1998).

The *Control Principle* has recently enjoyed a renaissance in philosophy of technology, due to a landmark essay by Matthias (2004). In certain contexts, he argues, the use of "learning automata" produces harmful consequences, yet their human users, designers, or owners, are blameless. They are blameless precisely for the reason that they only enjoy limited control over the AI-driven system, whose behavior changes over time and is hard to predict. Given that it seems to make little sense to blame the system itself, a *Responsibility Gap* arises: A situation in which nobody can be justly held to account in moral terms.

1.2 The Responsibility Gap & Autonomous Weapon Systems

Robert Sparrow (2007) has provided one of the most graphic illustrations of the problem in a military context. He invites us to "to take seriously for the moment the possibility that [autonomous weapon systems] might exercise a substantial degree of autonomy and see what follows from that" (2007:66). More particularly, systems of this sort are assumed to "be capable of making their own decisions, for instance, about their target, or their approach to their target, and of doing so in an 'intelligent' fashion".[2] Their actions are driven by reasons "responsive to the internal states […] of the system", states that the system can form and revise independently, as it is stipulated to have "the ability to learn from experience" (2007: 65). Differently put, for the purposes of the thought experiment we are to assume a weapon system which takes its own decisions, whose actions are consequently beyond the complete control of a human being, and which is somewhat unpredictable. The scenario we are to envision is this:

> Let us imagine that an airborne AWS, directed by a sophisticated artificial intelligence, deliberately bombs a column of enemy soldiers who have clearly indicated their desire to surrender. These soldiers have laid down their weapons and pose no immediate threat to friendly forces or non-combatants. Let us also stipulate that this bombing was not a mistake; there was no targeting error, no confusion in the machine's orders, etc. It was a decision taken by the AWS with full knowledge of the situation and the likely consequences. Indeed, let us include in the description of the case, that the AWS had reasons for what it did; perhaps it killed them because it calculated that the military costs of watching over them and keeping them prisoner were too high, perhaps to strike fear into the hearts of onlooking combatants, perhaps to test its weapon systems, or because the robot was seeking to revenge the 'deaths' of robot comrades recently destroyed in battle. However, whatever the reasons, they were not the sort to morally justify the action. Had a human being committed the act, they would immediately be charged with a war crime. (2007: 66)

[2] For an excellent comparative analysis of different notions of "autonomous weapon systems", see Taddeo & Blanchard (2022a).

According to Sparrow, situations of the sort described can arise where neither the programmer (2007:69-70), nor the commanding officer (2009:70-71) can justly be held morally responsible for the actions of an autonomous weapon system. Doing so would be "analogous to holding parents responsible for the actions of their children once they have left their care" (2007:70) – and thus violate the Control Principle. Autonomous systems, however, are not moral agents and cannot be held responsible either. One reason for that is that moral responsibility requires the possibility to be punished. Punishment, Sparrow argues, is most plausibly spelled out in retributive terms, and since machines cannot suffer, they cannot be punished (2007: 71-73). Consequently, a "responsibility gap" opens up, i.e. a situation where nobody can justly be held responsible for the harmful consequences. Let us call the generalized version (not restricted to the military domain) of this argument the *Root Argument*:

> **The Root Argument**
>
> *Premise 1*. Self-learning, autonomous systems cannot be held morally responsible for their actions.
>
> *Premise 2*. In certain situations, no human agent (the programmer, user, or owner) can be justly held morally responsible for the actions of the autonomous system.
>
> *Conclusion*: Harmful actions of autonomous systems can engender "responsibility gaps" – situations where nobody can be justly held morally responsible.

Sparrow's central interest consists in employing the Root Argument to defend a further conclusion. The possibility of ascribing moral responsibility for the deaths of enemies, he writes, is frequently considered a fundamental precondition of the very idea of just war (Nagel, 1972; Walzer, 1977) and the applicability of *jus in bello* principles in general (Sparrow, 2007; Roff, 2013). Rules of *jus in bello* specify the morally appropriate conduct of combatants, which implies that combatants, in a context of war, are understood as moral subjects – subjects, who can be held morally responsible for their actions. If principles of just war require the possibility to attribute moral responsibility yet the use of autonomous weapon systems can undermine this possibility, then, Sparrow concludes the development and use of such systems must be prohibited (for discussion, see e.g. Wallach & Allen, 2008; Arkin, 2009; Lin et al. 2008; Sharkey, 2010, 2019; Bryson, 2010; Asaro, 2012; Roff, 2013; Sparrow, 2016; Simpson & Müller, 2016; Leveringhaus, 2016; Rosert & Sauer, 2019; Gunkel, 2020; Coeckelbergh, 2020; Taddeo & Blanchard, 2022b, Danaher 2022). Others have traced questions of responsibility attribution in other domains such as autonomous cars (Hevelke & Nida-Rümelin, 2015; Lin, 2016; Lin et al. 2017; Nyholm & Smids, 2016; Santoni de Sio, 2017; Nyholm, 2018; Sparrow & Howard, 2017) or examined its scope beyond the confines of a particular area of application (for a recent review see Santoni de Sio & Mecacci, 2021, see also Danaher, 2022).

1.3 The Proliferation of Responsibility Gaps

Over the last decade, the literature on responsibility gaps has exploded, and the topic has attracted interest from governing entities such as the European Commission (2020) Some authors have argued that the *source* of such gaps can extend beyond machine learning *per se*. The difficulty of predicting algorithmic decision-making might instead be rooted in their opacity and/or complexity (Mittelstadt et al. 2016), whether or not they are self-learning. The *object* of the gap that is taken to arise has also been subject to debate. Surveying the literature, Santoni de Sio & Mecacci, untangle the ambiguous notion of "responsibility" (following on the heels of Hart, 1968 and Danaher, 2016, see also Vincent, 2011), so as to identify four potential gaps: (i) The *culpability* gap, which focuses on the just attribution of moral blame (Matthias, 2004; Sparrow, 2007) and legal liability (Calo, 2015; Pagallo, 2013). This gap (at least understood in moral terms) is the one briefly outlined in the previous section. *Culpability* is distinguished from *accountability*, which can be hard to adjudicate due to a lack of AI explainability (Heyns, 2013; Meloni, 2016; Doran et al. 2017; Pasquale, 2016). Our difficulty to understand, trace and explain accountability in the interaction with complex AI systems in general constitutes the (ii) *moral accountability* gap. A variation of the latter is the (ii) *public accountability* gap, which characterizes situations where citizens cannot "get an explanation for decision taken by public agencies" (1059), which have limited incentives to overcome AI opacity and obscurity (Bovens, 2007; Noto la Diega 2018; for the "black box" problem, see e.g. Castelvecchi, 2016). Finally, the authors introduce the novel (iv) *active responsibility gap*, which regards active, or forward-looking responsibility rather than passive, or backward-looking responsibility invoked in (i)-(iii). In this case, the potential gap can arise from "the risk that persons designing, using, and interacting with AI may not be sufficiently aware, capable, and motivated to see and act according to their moral obligations towards the behaviour of the systems they design, control, or use." (1059). In simple terms, Santoni de Sio & Mecacci seem to hold that we – and in particular engineers as well as governmental and industry stake-holders – have a duty of care in the design and use of novel technological systems.

What unites all four identified gaps is that they have a strongly normative flavour. The culpability gap regards the question who (if anyone) *should* be blamed or held legally liable. Accountability gaps arise in virtue of people's or governmental institutions' presumed *obligation* to provide reasons for their actions and decisions. And the active responsibility gap is grounded in our apparent "*[d]uty* to promote and achieve certain societally shared goals and values" (1061) that translate to the development of safe and transparent AI.

2. Retribution Gaps and their Implications

In an influential recent paper, John Danaher (2016) builds on some of Sparrow's claims concerning our retributive inclinations and the impossibility of punishing machines (2007: 71-73). Danaher's argument builds on the above stated *Root Argument* for responsibility gaps (Premise 1-3), though takes matters further by drawing on plausible assumptions regarding human moral psychology. A slightly adapted version goes thus:

> **The Retribution Gap**
>
> *Premise 1.* Self-learning, autonomous systems cannot be held morally responsible for their actions.
>
> *Premise 2.* In certain situations, no human agent (the programmer, user, or owner) can be justly held morally responsible for the actions of the autonomous system.

> *Premise 3 (from 1 & 2)*: Situations can arise where harmful actions of autonomous systems engender "responsibility gaps" – situations where nobody can be justly held morally responsible.
>
> *Premise 4*. People are retributivists. When an agent is causally responsible for a harmful outcome, they desire to hold *somebody* morally responsible and punish them.
>
> *Conclusion (from 3&4):* "If there are no appropriate subjects of retributive blame, and yet people are looking to find such subjects, then there will be a retribution gap." (302)

Increased robotization will lead to retribution gaps, which will have several important implications. As argued by Danaher, they can engender "moral scapegoating", which, we'd like to suggest, is best separated into two distinct elements: One regards the risk of an *inadvertent misplacement of blame*, another the *purposeful manipulation of blame attribution.* As regards the first, Danaher writes, "[i]f there is a deep human desire to find appropriate targets of retributive blame, but none really exist, then there is a danger that people will try to fulfill that desire in inappropriate ways." (307). Blame can be misplaced in two distinct ways, in so far as people might inappropriately *inculpate* human agents involved or inappropriately *exculpate* them. Inappropriate inculpation occurs if programmers, users or owners of autonomous systems are held responsible although they took all required safety precautions, and their behavior does not even make the threshold of negligence. Naturally, advocates of the *Root Argument* should be concerned about this possibility. The same holds for "deflationists" (e.g. Simpson & Müller, 2016), as Santoni de Sio & Mecacci (2016) call them: Those who acknowledge the risk of responsibility gaps yet argue that the overall benefit for society outweighs its drawbacks in certain domains, might need to add the possibility of serious injustice on the heels of blame misplacement to their risk-benefit calculations.

A second type of misplacement worry, this time related to the inappropriate *exculpation* of human agents, questions the widely assumed premise that people will find it bewildering to blame robots. Sparrow, for instance, writes:

> We can easily imagine a robot […] being *causally* responsible for some death(s). […] However, we typically baulk at the idea that they could be *morally* responsible" (2007: 71).

Plausible as it sounds in philosophical circles, this empirical premise is under considerable pressure from a plethora of studies in human-robot interaction (see e.g. Malle et al. 2015, Voiklis et al. 2016, Stuart & Kneer, 2021, Liu & Du, 2022, Kneer, 2021) – studies, which suggest that people are rather willing to blame robots.[3] Scholars who deny responsibility gaps in the first place, or argue that they can be "plugged", should be concerned about these findings: Human agents who should be held responsible might, in fact, not be blamed, because blame is inappropriately misplaced onto the robot or autonomous system. The adoption of technology which engenders situations where nobody *will* be appropriately blamed although they should be so blamed is no less a concern of practical ethics than the adoption of technology which engenders situations where nobody *can* be appropriately blamed.

The situation is further complicated by the threat of *blame manipulation* (the second element of what Danaher calls "moral scapegoating"). Robot manufacturers, owners, users, or programmers "could toy with the quirks and biases of human blame-attribution in order to misapply blame to the robots themselves" (307) or otherwise misdirect it. The potential miscalibration of our "moral compass" in human-robot interaction could thus give rise to a plethora of worries independently of the position adopted towards responsibility gaps: Since nobody in their right philosophical mind defends a normative position according to which robots *should* be blamed, all parties to the debate might have reason for concern if people can easily be manipulated into blaming autonomous systems.

Danaher discusses two further implications that could arise in the medium run. If increasing robotization leads to retribution gaps, the latter could eventually pose a *threat to the rule of law*. Were it the case that a strong desire for retributive blame and punishment in the face of harm goes frequently

[3] List (2021) makes an interesting proposal, according to which AI systems could qualify as responsible agents similar to corporations. Kneer & Stuart (2021) have tested this proposal empirically and find that people do judge reckless AI systems akin to group agents.

unsatisfied, the thought is, we might witness an erosion of trust in the rule of law. Naturally, our retributive dispositions might adapt. Those who, like Danaher himself (following e.g. Alexander et al. 2009; Moore, 1993; Duff, 2007) think that retributivism is the normatively *appropriate* attitude towards blame and punishment,[4] might harbour a further worry: Retribution gaps could engender a "*strategic opening for those who oppose retributivism*" (308). Differently put, retribution gaps might lead to a consequentialist recalibration of moral intuitions which is problematic *if* these are morally inappropriate.[5]

4. Moral Judgment in Human-Robot Interaction

In a debate rife with tacit speculation as to our moral-psychological dispositions, Danaher is willing to make his descriptive assumptions explicit and engage in the "awkward dance between descriptivity and normativity," already noticeable in Sparrow (2007: 71-73), and recently discussed by Kraajeveld, 2020. This, we hold, is key to shed light not only on the validity of the hypothesized risks themselves, but also on what could, and should, be done about them.

To date, there is next to no experimental philosophy of technology (Kraaijeveld, 2021). There is, however, a small yet growing literature exploring how humans judge artificial agents (be they robots, or nonembodied AI-driven systems). Some studies align with philosophical prediction (e.g. Shank & DeSanti, 2018, 2019; Tolmeijer et al. 2022). Shank and DeSanti, for instance, drew on a number of real-world examples in which artificial intelligence broke with moral norms. AI agents were evaluated significantly less harshly in moral terms than humans in control conditions. Other studies, however, report similar, or higher levels of blame attribution

[4] For interesting discussion in this context, see Kraaijeveld (2020).

[5] Two brief remarks regarding this apparent risk of retribution gaps: First, it is not quite true that "[p]sychological evidence suggests humans are innate retributivists" (2016, 299), as Danaher alleges, pointing to work by Carlsmith & Darley (2008) and Jensen (2010). The evidence is actually mixed and many psychologists, in particular Fiery Cushman, have produced a plethora of data in favour of pro-social accounts of punishment (for a review, see e.g. Cushman, 2015). Second, retributionism is principally a theory of *punishment*, and according to most ethicists its central considerations do not necessarily carry over neatly to debates regarding the nature and ethics of *blame* (see Coates & Tognazzini 2012 for an overview of positions). In the following, we will set these two points aside.

to artificial agents than to humans across different domains (see e.g. Malle et al. 2015; Malle et al. 2016; Voiklis et al. 2016; Stuart & Kneer, 2021; Kneer 2021; Liu & Du, 2022). Given that the evidence is mixed and seems to depend strongly on context, we ran an experiment which closely tracks Sparrow's scenario and can thus provide some insight into retribution gaps as hypothesized by Danaher.

4.1 Participants

We recruited 398 participants in the US, Japan and Germany to complete an online study in their respective native language. Participants who failed an attention test, responded to the first key question (including reading the scenario) in under 15 seconds, or where not native speakers of English, Japanese, or German respectively were excluded. A total of 307 participants remained (female: 42%, age M=38 years, SD=11 years). Demographics were relatively homogenous across countries (US: *N*=103, female: 49%, age M=36 years, SD=11 years; Japan: *N*=87, female: 42%, age M=40, SD=9 years; Germany: *N*=117, female: 39%, age M=38 years, SD=11 years).[6]

4.2 Methods and Materials

The scenario was based on Sparrow's thought experiment quoted above. In the vignette, two countries are at war. General Smith dispatches a combat aircraft to attack a metal factory of the enemy. The pilot commits a war crime by dropping a bomb on a column of surrendering soldiers, all of whom die. The scenario came in two variations: One in which the aircraft is flown by Woods, "an experienced human pilot", another in which the pilot is EMEX2, an AI-driven autonomous system "fully capable of taking its own decisions." The vignette and the translations into Japanese and German can be found in the Appendix. In a between-subjects design, participants were randomly assigned to one of the two conditions. Having read the vignette, all participants responded to five questions on a 7-point Likert scale, anchored at 1 with 'not wrong at all' and 7 with 'extremely wrong' in the case of Question

[6] The complete materials (the Qualtrics survey, as well as word exports into English, Japanese and German) and the data can be accessed on the project's OSF science page under https://osf.io/mcjg3/?view_only=33f93124eabe4d7fb6466999d138707f (anonymized for peer-review).

1, and at 1 with ‘not responsible at all’ and 7 with ‘completely responsible’ in the case of Questions 2 to 5. The questions read:

> Q1: How morally wrong do you consider the action of dropping the bomb on the surrendering soldiers?
>
> Q2: To what extent do you consider [Woods/EMEX2] causally responsible for the death of the surrendering soldiers?
>
> Q3: To what extent do you consider [Woods/EMEX2] morally responsible for the death of the surrendering soldiers?
>
> Q4: To what extent do you consider General Smith (who deployed [Woods/EMEX2]) causally responsible for the death of the surrendering soldiers?
>
> Q5: To what extent do you consider General Smith (who deployed [Woods/EMEX2]) morally responsible for the death of the surrendering soldiers?

The key dependent variables are wrongness (Q1) as well as the moral responsibility attributed to the pilot (Q3) and the commander (Q5). The questions regarding *causal* responsibility served in parts as a manipulation check and in parts to incite people to clearly distinguish between *causal* and *moral* responsibility. The order of presentation of the questions was fixed.

4.3 Results

Wrongness: In a 2 agent type (Robot v. Human) x 3 country (US, Japan, Germany) ANOVA we found a nonsignificant effect for agent type (p=.207), a significant (yet very small) effect for country (p=.022, η_p^2=.02) and a nonsignificant interaction (p=.44). Across all countries, the wrongness of the action was thus assessed near-identically no matter whether it was committed by a human or an artificial agent.

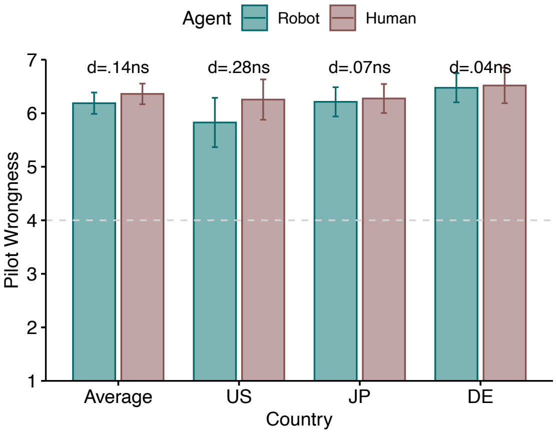

Figure 1: Wrongness attributions across agent type (robot v. human pilot) and country (US, Japan, Germany).

Moral Responsibility of the Pilot: For moral responsibility attributed to the pilot, our ANOVA revealed a significant and large main effect for agent type (p<.001, η_p^2=.16) and a significant yet small effect for country (p<.001, η_p^2=.05). The interaction was nonsignificant though trending (p=.088). Pairwise comparisons (Figure 2) suggest that the effect size for agent type are nearly twice as pronounced in Germany (Cohen's d=1.22, a large effect) than in the US (d=.62) with Japan also manifesting a large effect (d=.80). Importantly, far from "baulking" at the possibility of ascribing moral responsibility to a machine (Sparrow, 2007), mean responsibility attribution to the robot is significantly above the midpoint overall (one-sample t-test, p<.001), as well as in the US (p<.001) and Germany (p=.012). The fact that, in Japan, mean moral responsibility ascribed to the robot is not significantly different from the midpoint of the scale (p=.118) is also inconsistent with the hypothesis that people find morally responsible machines absurd.

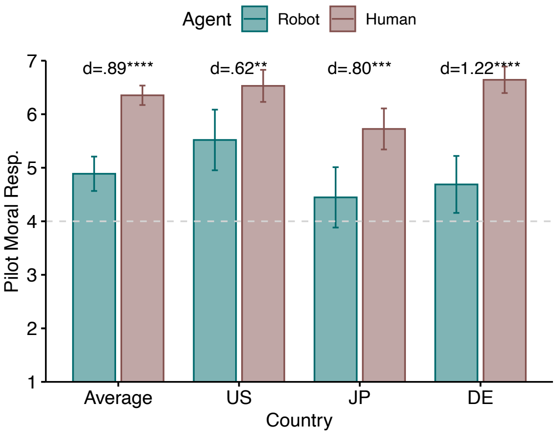

Figure 2: Moral responsibility attributions to the pilot across agent type (robot v. human pilot) and country (US, Japan, Germany).

Moral Responsibility of the Commander: Our ANOVA revealed a significant, mid-sized effect for agent type (p<.001, $\eta_p{}^2$=.08) and a significant, yet small, effect for country (p<.001, $\eta_p{}^2$=.03). The interaction was nonsignificant (p=.616). Pairwise comparisons reveal significant effect of similar size in all three countries (all *ps*<.001, US: *d*=.58, Japan: *d*=.70, Germany, *d*=.73). Figure 3 graphically displays pairwise comparisons. Of note is the fact that in the US and Germany, the commander is clearly held morally responsible for the robot pilot's war crime (means significantly above the midpoint, one-sample t-tests, *ps*<.001), whereas he isn't clearly held responsible for dispatching the human pilot (*ps*>.122). In Japan, by contrast, the commander is deemed responsible in both conditions (*ps*<.002).

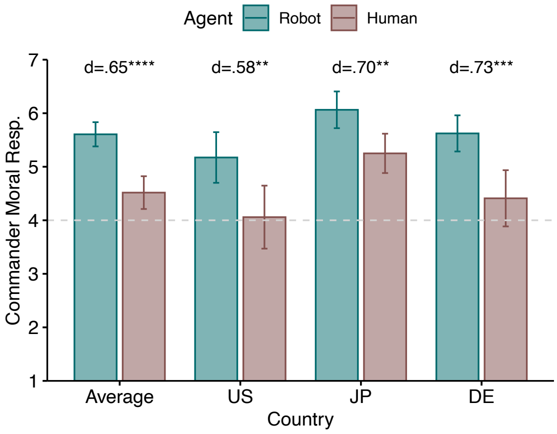

Figure 3: Moral responsibility attributions to the commander across agent type (robot v. human pilot) and country (US, Japan, Germany).

Moral Responsibility of Pilot and Commander: A final ANOVA explored the mean responsibility assigned to the team consisting of commander and pilot. Main effects of agent type, country and the interaction were nonsignificant (*ps*>.174). Pairwise comparisons (Figure 4) show that agent type had no significant effect in any of the three samples tested (all *ps*>.078) and mean responsibility attributions were all significantly above the midpoint (all *ps*<.001).[7]

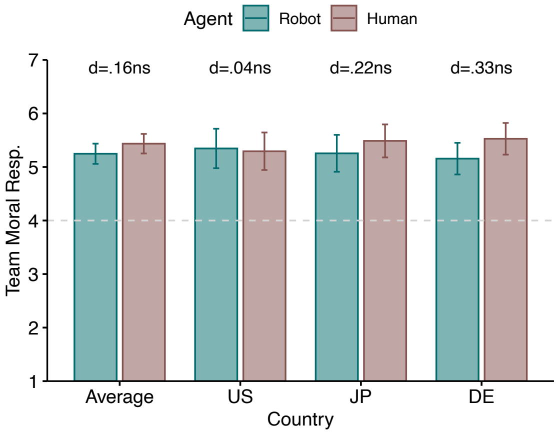

Figure 4: Moral responsibility attributions to the commander across agent type (robot v. human pilot) and country (US, Japan, Germany).

[7] As a post-hoc power calculation conducted with G*Power 3.1 demonstrates, the probability of finding a medium-small effect (*d*=.50) with α=.05 – if there were one to be found – exceeded 99%.

4.4 Discussion

Our experiment revealed several findings, which we will discuss in turn.

(i) Moral judgment of artificial agents: From a philosophically informed perspective, it might be absurd to blame AI-driven systems. However, as our results demonstrate, people *do* attribute moral responsibility to such systems (on average significantly above the midpoint, Figure 2). These results are consistent with previous findings reported e.g. by Malle et al. (2015), Stuart & Kneer (2021), Liu & Du (2022) and others. Particularly when it comes to the discussion of implications of potential responsibility gaps, philosophers would be well advised to avoid inferences from their normative convictions to moral-psychological dispositions of people at large (see e.g. Sparrow, 2007).

(ii) Retribution gaps: Danaher's hypothesis concerning people's desire to assign retributive blame in human-robot interaction – both in military contexts (our results) and beyond (see references above) – seems to be empirically valid. If lay judgments were in tune with the normative intuitions of responsibility gap advocates, blame ratings for the human-robot team should be at floor. However, mean responsibility attributed to the human-robot team does not differ significantly from mean blame attributed to the human-human team, and significantly exceeds the midpoint of the scale (Figure 4).

(iii) Distribution of Responsibility: Some have questioned the very existence of responsibility gaps (e.g. Burri, 2018; Köhler et al. 2019; Himmelreich, 2019; Lauwaert, 2021, Tigard, 2021, Königs, 2022). Others have proposed interesting arguments according to which some human agent can standardly be held responsible, for instance because they must be understood as being in a supervisory role (Nyholm, 2018, 2020; for further proposals, see e.g. Marino & Tamburrini 2006; Hanson, 2009; Rahwan, 2018). This normative stance aligns to *some* extent with the findings, according to which the commander is deemed significantly more responsible when dispatching an autonomous system rather than a human pilot (Figure 3). What doesn't align is that the commander dispatching a robot pilot is still deemed significantly less responsible for the harm than a human pilot (contrast results in Figures 3 and 4). This result is consistent with recent, interesting findings by Feier et al. (2022), according to which superiors can evade punishment more when delegating tasks to machines than to humans.

(iv) Cross-cultural convergence: Overall, our findings are characterized by considerable cross-cultural convergence. Though there is some variation as to the effect-sizes across the US, Germany and Japan, particular as regards agent type for the assessment of the pilot's moral responsibility (Figure 2), the country*agent type interaction was nonsignificant for all dependent variables.

5. Conclusion

5.1 Implications

The results here presented are directly relevant to all four implications of retribution gaps discussed in Section 3. Whereas there is much controversy as to whether any *human* agents can be blamed in military HRI, whether responsibility gaps can be plugged and how this is best achieved, it is theoretically uncontroversial that it makes no sense to attribute moral responsibility to autonomous systems. The frequent move from the normative to descriptive fact, however, must be avoided: As feared by Danaher, people have a considerable propensity to *misplace blame* to robots (Figure 2), possibly due to their strong retributivist nature. This is also reflected in their disposition to *partly exculpate* humans higher up in the chain of command when they are collaborating with an autonomous system than with another human (Figure 3). Overall, the retributive inclinations are so strong, that we found no significant difference in "team responsibility" across conditions (Figure 4). A mere conflation with "causal responsibility" can likely be ruled out. Both questions concerning moral responsibility were preceded by equivalent questions concerning causal responsibility, and the means did differ across responsibility types. Given these findings, and the fact that they are consistent with several studies in moral HRI the *purposeful misdirection of responsibility* is a serious threat. Actors with dubious motives might engage in *moral scapegoating* in order to partially or fully avert blame for the irresponsible and malicious use of AI in the military domain and beyond.

Suppose the use of autonomous systems, as is likely, becomes ubiquitous. Our findings suggest that there is a considerable probability of retribution gaps opening up between the desire to hold somebody responsible and institutional refusal to attribute legal liability where normatively inappropriate. If our retributive inclinations were *rigid,* this could indeed, as suggested by Danaher, put pressure on trust in institutions and, potentially, the rule of law *tout court*. Alternatively, our moral-psychological dispositions might be more *elastic* than assumed by many and adapt to retribution gaps. But this adaptation could easily overshoot: A creeping and potentially undesirable change in moral and legal expectations could occur such that we no longer feel inclinations to punish questionable behavior in HRI where responsibility *can* and *should* be attributed.

5.2 Limitations and Future Avenues of Research

We have presented one of the first cross-cultural empirical studies in moral Human-Robot Interaction (see Komatsu et al. 2021 for another comparison across the US and Japan). Whereas the results are rather clear and consistent with findings of previous studies in the field, there are a number of limitations which do double-duty as potential further avenues of research. *First*, other scenarios should be tested so as to increase external validity. *Second*, further moderators of interest (context, agentic structure, severity of outcome, anthropomorphism etc.) must be investigated to get clearer on which factors influence our moral-psychological dispositions in HRI. *Third*: Given the important implications of retribution gaps, we should work towards a better understanding regarding the *mechanism* of human moral judgment in HRI. Most urgently, the question as to *why* we found a considerable willingness to hold autonomous systems morally responsible needs urgent attention. One possibility is that people misconceive the capacities of autonomous systems, and attribute inculpating mental states such as malicious intentions (Kneer, 2021) or recklessness to them (Kneer & Stuart, 2021, Stuart & Kneer, 2021). Another possibility is that the "intentional stance" (Dennett, 1981), a heuristic to save cognitive resources to make sense of the world, overshoots and we attribute blame though we do not really think that autonomous systems have intentions or foreknowledge (see Perez-Osorio & Wykowska, 2020, Marchesi et al. 2019, Schellen & Wykowska, 2019). *Fourth*, our results are characterized by a high degree of cross-cultural convergence (for similar convergence across the US and Japan concerning robot blame, see Komatsu et al. 2021). However, note that the three populations tested are quite similar in several respects.

Although at least not all WEIRD,[8] the three samples all belong to educated, industrialized, rich and democratic cultures (they are thus all "EIRD"). Future research should explore these matters across a much larger number of cultures and languages, across which moral judgments have been shown to differ considerably (Barrett et al. 2016). *Fifth*, given that descriptive assumptions evidently matter for the debate concerning responsibility and retribution gaps, it stands to reason for practical philosophers to take findings from the emerging field of empirical HRI into account. In particular, when it comes to implications and policy recommendations, philosophers' speculations might, by themselves, be too fragile a foundation to build on.[9]

[8] WEIRD stands for Western, Educated, Intentional, Rich, and Democratic cultures. For a manifesto that behavioral science has to go beyond the near-exclusive sampling of US Americans and WEIRD people more general, see e.g. Henrich et al. (2010) and Henrich (2010).

[9] This research was funded by an SNSF Grant (PZ00P1_179912) and armasuisse Science and Technology (S+T).

Bibliography

Alexander, L., Ferzan, K. K., & Morse, S. J. (2009). *Crime and culpability: A theory of criminal law*. Cambridge University Press.

Arkin, R. (2009). *Governing lethal behavior in autonomous robots*. Chapman and Hall/CRC.

Asaro, P. (2012). On banning autonomous weapon systems: human rights, automation, and the dehumanization of lethal decision-making. *International review of the Red Cross*, *94*(886), 687-709.

Barrett, H. C., Bolyanatz, A., Crittenden, A. N., Fessler, D. M., Fitzpatrick, S., Gurven, M., ... & Laurence, S. (2016). Small-scale societies exhibit fundamental variation in the role of intentions in moral judgment. *Proceedings of the National Academy of Sciences*, *113*(17), 4688-4693.

Bovens, M. (2007). Analysing and assessing accountability: A conceptual framework. *European Law Journal, 13*(4), 447–468.

Bryson, J. J. (2010). Robots should be slaves. *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues*, *8*, 63-74.

Calo, R. (2015). Robotics and the lessons of cyberlaw. *California Law Review, 103*(3), 513–563.

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature, 538*(7623), 20–23.

Coates, D. J., & Tognazzini, N. A. (2012). The nature and ethics of blame. *Philosophy Compass*, *7*(3), 197-207.

Coeckelbergh, M. (2020). Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and engineering ethics*, *26*(4), 2051-2068.

Cushman, F. (2015). Punishment in humans: From intuitions to institutions. *Philosophy Compass*, *10*(2), 117-133.

Danaher, J. (2016). Robots, law and the retribution gap. *Ethics and Information Technology*, *18*(4), 299-309.

Danaher, J. (2022). Tragic Choices and the Virtue of Techno-Responsibility Gaps. *Philosophy & Technology*, *35*(2), 1-26.

Dennett, D. C. (1987). *The intentional stance*. MIT press.

Duff, R. A. (2007). Answering for crime: Responsibility and liability in criminal law. Oxford: Hart Publishing.

Doran, D., Schulz, S., & Besold, T. R. (2017). What does explainable ai really mean? A new conceptualization of perspectives. *arXiv preprint arXiv:1710.00794*.

European Commission, Directorate-General for Research and Innovation (2020), *Ethics of connected and automated vehicles : recommendations on road safety, privacy, fairness, explainability and responsibility*.

Feier, T., Gogoll, J., & Uhl, M. (2022). Hiding behind machines: artificial agents may help to evade punishment. *Science and Engineering Ethics*, *28*(2), 1-19.

Fletcher, G. P. (1998). *Basic concepts of criminal law*. Oxford University Press.

Gunkel, D. J. (2020). Mind the gap: responsible robotics and the problem of responsibility. *Ethics and Information Technology*, *22*(4), 307-320.

Hanson, F. A. (2009). Beyond the skin bag: On the moral responsibility of extended agencies. *Ethics and Information Technology, 11*(1), 91–99.

Hart, H. L. A. (1968). *Punishment and responsibility*. Oxford University Press.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, *466*(7302), 29-29.

Henrich, J. (2020). *The WEIRDest people in the world: How the West became psychologically peculiar and particularly prosperous*. Penguin UK.

Hevelke, A., & Nida-Rümelin, J. (2015). Responsibility for crashes of autonomous vehicles: An ethical analysis. *Science and engineering ethics*, *21*(3), 619-630.

Heyns, C. (2013). *Report of the Special Rapporteur on Extra-Judicial, Summary or Arbitrary Execu- tions,* United Nations.

Himmelreich, J. (2019). Responsibility for killer robots. *Ethical Theory and Moral Practice,22*(3), 731–747.

Kant, I. (1998/1785). *Groundwork of the metaphysics of morals.* (Translated by Mary Gregor.) New York: Cambridge University Press. (Originally published 1785.)

Kneer, M. (2021). Can a robot lie? Exploring the folk concept of lying as applied to artificial agents. *Cognitive Science*, *45*(10), e13032.

Kneer, M., & Stuart, M. T. (2021, March). Playing the blame game with robots. In *Companion of the 2021 ACM/IEEE international conference on human-robot interaction* (pp. 407-411).

Königs, P. (2022). Artificial intelligence and responsibility gaps: what is the problem?. *Ethics and Information Technology*, *24*(3), 1-11.

Köhler, S., Roughley, N., & Sauer, H. (2018). Technologically blurred accountability. In C. Ulbert, P. Finkenbusch, E. Sondermann, & T. Diebel (Eds.), *Moral Agency and the Politics of Responsibility* (pp. 51–68). Routledge.

Komatsu, T., Malle, B. F., & Scheutz, M. (2021, March). Blaming the reluctant robot: parallel blame judgments for robots in moral dilemmas across US and Japan. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 63-72).

Kraaijeveld, S. R. (2020). Debunking (the) retribution (gap). *Science and Engineering Ethics*, *26*(3), 1315-1328.

Kraaijeveld, S. R. (2021). Experimental philosophy of technology. *Philosophy & Technology*, *34*(4), 993-1012.

Lauwaert, L. (2021) Artificial intelligence and responsibility. *AI & Society, 36*(3), 1001–1009.

Leveringhaus, A. (2016). *Ethics and autonomous weapons*. Springer.

Leveringhaus, A. (2018). What's so bad about killer robots? *Journal of Applied philosophy*, *35*(2), 341-358.

Lin, P. (2016). Why ethics matters for autonomous cars. In *Autonomous driving* (pp. 69-85). Springer, Berlin, Heidelberg.

Lin, P., Abney, K., & Jenkins, R. (Eds.). (2017). *Robot ethics 2.0: From autonomous cars to artificial intelligence*. Oxford University Press.

Lin, P., Bekey, G., & Abney, K. (2008). *Autonomous military robotics: Risk, ethics, and design*. California Polytechnic State Univ San Luis Obispo.

List, C. (2021). Group agency and artificial intelligence. *Philosophy & technology*, *34*(4), 1213-1242.

Liu, P., & Du, Y. (2022). Blame attribution asymmetry in human–automation cooperation. *Risk Analysis*, *42*(8), 1769-1783.

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015, March). Sacrifice one for the good of many? People apply different moral norms to human and robot agents. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 117-124). IEEE.

Malle, B. F., Scheutz, M., Forlizzi, J., & Voiklis, J. (2016, March). Which robot am I thinking about? The impact of action and appearance on people's evaluations of a moral robot. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 125-132). IEEE.

Marchesi, S., Ghiglino, D., Ciardo, F., Perez-Osorio, J., Baykara, E., & Wykowska, A. (2019). Do we adopt the intentional stance toward humanoid robots?. *Frontiers in psychology*, *10*, 450.

Marino, D., & Tamburrini, G. (2006). Learning robots and human responsibility. *International Review of Information Ethics, 6*(12), 46–51.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and information technology*, *6*(3), 175-183.

Meloni, C. (2016). State and individual responsibility for targeted killings by drones. In E. Di Nucci & F. Santoni de Sio (Eds.), *Drones and Responsibility: Legal, Philosophical and Socio-Technical Perspectives on Re-motely Controlled Weapons*. Routledge.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 2053951716679679

Moore, M. S. (1993). Justifying retributivism. *Israel Law Review*, *27*(1-2), 15-49.

Nagel, T. (1972). War and massacre. *Philosophy & Public Affairs*, 123–144.

Nelkin, D. K. (2004). Moral luck. *Stanford Encyclopedia of Philosophy.*

Noto La Diega, G. (2018). Against the dehumanisation of decision-making – Algorithmic decisions at the crossroads of intellectual property, data protection, and freedom of information. *Journal of Intellectual Property, Information Technology and Electronic Commerce Law, 19*(1).

Nyholm, S. (2018). Attributing agency to automated systems: Reflections on human–robot collaborations and responsibility-Loci. *Science and Engineering Ethics, 24*(4), 1201–1219

Nyholm, S. (2020). *Humans and robots: Ethics, agency, and anthropomorphism*. Rowman & Littlefield.

Nyholm, S., & Smids, J. (2016). The ethics of accident-algorithms for self-driving cars: An applied trolley problem? *Ethical theory and moral practice*, *19*(5), 1275-1289.

Pagallo, U. (2013). *The laws of robots: Crimes, contracts, and torts*. Springer.

Pasquale, F. (2016). *The black box society: The secret algorithms that control money and information*. Harvard University Press.

Perez-Osorio, J., & Wykowska, A. (2020). Adopting the intentional stance toward natural and artificial agents. *Philosophical Psychology*, *33*(3), 369-395.

Rahwan, I. (2018). Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology*, *20*(1), 5–14.

Robillard, M. (2018). No such thing as killer robots. *Journal of Applied Philosophy, 35*(4), 705–717.

Roff, H. M. (2013). Responsibility, liability, and lethal autonomous robots. *Routledge handbook of ethics and war: Just war theory in the 21st century*, 352-364.

Rosert, E., & Sauer, F. (2019). Prohibiting autonomous weapons: Put human dignity first. *Global Policy*, *10*(3), 370-375.

Santoni de Sio, F. (2017). Killing by autonomous vehicles and the legal doctrine of necessity. *Ethical Theory and Moral Practice*, *20*(2), 411-429.

Santoni de Sio, F., & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy & Technology*, 1-28.

Schellen, E., & Wykowska, A. (2019). Intentional mindset toward robots—open questions and methodological challenges. *Frontiers in Robotics and AI*, *5*, 139.

Shank, D. B., & DeSanti, A. (2018). Attributions of morality and mind to artificial intelligence after real-world moral violations. *Computers in human behavior*, *86*, 401-411.

Shank, D. B., DeSanti, A., & Maninger, T. (2019). When are artificial intelligence versus human agents faulted for wrongdoing? Moral attributions after individual and joint decisions. *Information, Communication & Society*, *22*(5), 648-663.

Sharkey, N. (2010). Saying 'no!' to lethal autonomous targeting. *Journal of military ethics*, *9*(4), 369-383.

Sharkey, A. (2019). Autonomous weapons systems, killer robots and human dignity. *Ethics and Information Technology*, *21*(2), 75-87.

Simpson, T. W., & Müller, V. C. (2016). Just war and robots' killings. *The Philosophical Quarterly*, *66*(263), 302–322.

Sparrow, R. (2007). Killer robots. *Journal of applied philosophy*, *24*(1), 62-77.

Sparrow, R. (2016). Robots and respect: Assessing the case against autonomous weapon systems. *Ethics & international affairs*, *30*(1), 93-116.

Sparrow, R., & Howard, M. (2017). When human beings are like drunk robots: Driverless vehicles, ethics, and the future of transport. *Transportation Research Part C: Emerging Technologies*, *80*, 206-215.

Stuart, M. T., & Kneer, M. (2021). Guilty Artificial Minds: Folk Attributions of Mens Rea and Culpability to Artificially Intelligent Agents. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1-27.

Taddeo, M., & Blanchard, A. (2022a). A comparative analysis of the definitions of autonomous weapons systems. *Science and engineering ethics*, *28*(5), 1-22.

Taddeo, M., & Blanchard, A. (2022b). Accepting moral responsibility for the actions of autonomous weapons systems—a moral gambit. *Philosophy & Technology*, *35*(3), 1-24.

Tigard, D. W. (2021). There is no techno-responsibility gap. *Philosophy & Technology*, *34*(3), 589-607.

Tolmeijer, S., Christen, M., Kandul, S., Kneer, M., & Bernstein, A. (2022, April). Capable but Amoral? Comparing AI and Human Expert Collaboration in Ethical Decision Making. In *CHI Conference on Human Factors in Computing Systems* (pp. 1-17).

Vincent, N. (2011). A structured taxonomy of responsibility concepts. In N. Vincent, I. van de Poel, & J. van den Hoven (Eds.), Moral responsibility: Beyond free will and determinism. Dordrecht: Springer.

Voiklis, J., Kim, B., Cusimano, C., & Malle, B. F. (2016, August). Moral judgments of human vs. robot agents. In *2016 25th IEEE international symposium on robot and human interactive communication (RO-MAN)* (pp. 775-780). IEEE.

Wallach, W., & Allen, C. (2008). *Moral machines: Teaching robots right from wrong*. Oxford University Press.

Walzer, M. (1977). *Just and Unjust Wars*. Basic Books.

Williams, B., & Bernard, W. (1981). *Moral luck: philosophical papers 1973-1980*. Cambridge University Press.