Social Robots with AI: Prospects, Risks, and Responsible Methods
J. Seibt et al. (Eds.)
IOS Press, 2025

doi:10.3233/FAIA241510

"Responsibility" Plus "Gap" Equals "Problem"

Marc CHAMPAGNE[a,1]
[a] *Philosophy, Kwantlen Polytechnic University, Canada*
ORCiD ID: https://orcid.org/0000-0002-9841-7538

Abstract. Peter Königs recently argued that, while autonomous robots generate responsibility gaps, such gaps need not be considered problematic. I argue that Königs' compromise dissolves under analysis since, on a proper understanding of what "responsibility" is and what "gap" (metaphorically) means, their joint endorsement must repel an attitude of indifference. So, just as "calamities that happen but don't bother anyone" makes no sense, the idea of "responsibility gaps that exist but leave citizens and ethicists unmoved" makes no sense.

Keywords. Ethics, responsibility, autonomous robots

1. Introduction

We have not yet built or dealt with robots capable of making their own decisions. But, if and when such robots arrive, our usual moral practices will come up short. This is in part because robotic autonomy means giving up human control. An autonomous robot may act badly, not because of any malfunction or malicious programming, but simply because it decides to. Until now, only mature adults were credited with this ability, so we always knew who to hold responsible, at least in principle. Yet, given that a programmer has no say in a robot's decisions and robots feel no pain, our desire to blame lacks a suitable target. Matthias [1] called this the responsibility gap.

Sparrow [2] argued that this responsibility gap is serious enough to justify a ban on the creation and deployment of autonomous robots. Arkin [3] countered that such robots could be more moral than us. Champagne and Tonkens [4] hold that a human can voluntarily accept blame in advance for a robot's acts. Kiener [5] contends that a human can also do this after the fact. Burri [6] dismisses suggestions that robots could reach autonomy. Gunkel [7] invites us to regard robots as moral patients. Gogoshin [8] invites us to regard robots as moral agents. Søvik [9] says that autonomous robots cannot feel but may still be blamed. Tigard [10] argues that the problem prompting these various responses does not even exist.

These debates are not settled, so more stances continue to be built. Recently, Peter Königs [11] ventured onto new terrain by maintaining that the responsibility gap exists but is not problematic. Call this Königs' compromise. By all accounts, it represents a novel position. Yet, like any stance, we can ask: is it tenable?

Königs grants that, in some cases, autonomous robots generate responsibility gaps—since we must credit them with bad actions yet can never blame them [12]. He

[1]Corresponding Author: Marc Champagne; email: marc.champagne@kpu.ca.

summarizes this standard understanding well: "[T]he idea is that it is the system's autonomy […] that cancels people's responsibility. At the same time, the machines are not deemed sophisticated or 'agent-like' enough to be themselves possible bearers of responsibility" [11: 3]. Autonomous robots *decide* like human adults yet do not *feel* like human adults, so the negative consequences of their decisions cannot boomerang back to negatively affect them.

Most people regard this worst-of-both-worlds combination as worth avoiding. However, Königs claims that we can recognize the presence of responsibility gaps while feeling no pressure to seal or bridge them. As he puts it, if and when responsibility gaps "do occur, we need not be too concerned about them" [11: 2]. So, whereas Sparrow [2] took responsibility gaps to speak against the construction of (potentially lethal) autonomous robots, Königs seeks "to present a more optimistic view on artificial intelligence by dispelling such concerns about responsibility gaps" [11: 1].

Although I have no beef with techno-optimism [13], I will argue that Königs' compromise is untenable. While all participants to the debates value options and would celebrate the addition of a new possibility, the feasibility of a stance depends on the concepts and commitments it combines, since these constrain what can and cannot be truthfully said or inferred. I submit that, on a proper understanding of what "responsibility" is and what "gap" (metaphorically) means, their joint endorsement must repel an attitude of indifference.

2. Argument

Königs presents two challenges to those who regard responsibility gaps as problematic. The first challenge is to explain why robotic autonomy creates gaps in moral responsibility. The second challenge is to "explain why the existence of such gaps is problematic" [11: 2]. Most would say that Matthias [1] has already amply met the first challenge. Königs, however, says that "Matthias' original characterization of responsibility gaps is, in my view, rather unclear" [11: 2]. Since Königs never explains why Matthias' many concrete examples failed to live up to his expectations, Königs' basis for challenging a settled matter is, in my view, rather unclear. I will thus focus on Königs' second challenge. This must be feasible, because Königs also skips his first challenge when he grants that "such gaps may indeed arise" [11: 5].

Königs' compromise involves a decision to "not categorically deny the existence of responsibility gaps" [11: 2]. A distinction can indeed be made between existing and being problematic. It is fully possible, for example, for a hurricane to occur at sea in a way that is not (morally or practically) problematic for anyone. Yet, what happens when the analysis switches from "hurricane" to, say, "calamity"? Can a calamity exist yet fail to be problematic? The analytic-synthetic distinction may be blurry, but most would agree that a calamity involves loss and suffering (by one or more people). Hence, if some particular event x in the world satisfies the description "…is a calamity" ($\exists x \mathrm{C}x$), then this true existential claim guarantees that the same x is best described as worth avoiding ($\exists x \mathrm{C}x \rightarrow \exists x \mathrm{A}x$). A similar analytic entailment applies, I contend, to the responsibility gap.

Responsibility gaps are, by their nature, undesirable. Just as "calamities that happen but don't bother anyone" makes no sense, the idea of "responsibility gaps that exist but leave citizens and ethicists unmoved" (my expression) makes no sense.

If this conceptual/logical analysis and entailment is correct, then those concerned with the moral conundrum(s) posed by autonomous robots are under no special argumentative burden to "explain why the existence of such gaps is problematic" [11: 2]. An ontological acknowledgement of responsibility gaps is *by that very fact* an acknowledgement of their problematic nature.

While an outright denial of responsibility gaps (like the stance taken by Tigard [10]) would be logically consistent, the mixed stance taken by Königs is not. This, at any rate, is the gist of my concern with his compromise. Let me now go over my core argument more slowly, by unpacking the key terms "responsibility," "gap," and "problem" (in that order).

3. What "responsibility" means

The word "responsibility" figures in our evaluative lexicon, if anything does. Responsibility allows us (among other things) to track who did what and had a capacity to do otherwise. Insofar as this tracking enables judgment and judgment serves a vital welfare-promoting function, responsibility is a good. Bad events will invariably happen, but we try to reduce bad events stemming from human sources by justly directing blame, liability, punishments, and so on.

Like the law of supply and demand, when you dim the negative consequences of causing harm, you increase the likelihood that harm will be caused. This insight about the practical benefits of moral responsibility is as old as Plato's Ring of Gyges fable. The problem, however, is that our usual ethical resources fail to move or affect autonomous robots which decide to play hardball. Without a hint of mental or physical pain in the mix, there is no real way to deter such a robot from acting badly. Why should they care that others care? In this respect, autonomous social robots resemble sociopaths. They might act well some or most of the time, but given that they do not really care about our usual responsibility practices, our good fortune is at the mercy of their whim.

Having sociopaths in our midst is the price we pay for allowing the lottery of births to play itself out. Autonomous robots, by contrast, are fully avoidable—we spent almost all of human history without them. So, when harms are caused by such machines, we might be tempted to blame those who built them. The problem is that, owing to machine learning, "[t]he rules by which [such machines] act are not fixed during the production process, but can be changed during the operation of the machine, *by the machine itself*" [1: 177; emphasis in original]. This leaves us in a bind. We cannot blame a robot because it can't really care, but neither can we blame programmers because they had no say in the robot's particular decision(s).

This is cause for concern. We carefully praise and blame in the hope that people will make good choices and reason will prevail. It is hard to see how one could both 1) acknowledge that a new class of agents is unaffected by these usual incentives and constraints and that 2) all of this is unproblematic. If the ability to assign responsibility is good—as I am sure Königs would grant—then the inability to assign responsibility is bad.

Despite this, "the degree in which our society depends on the use of [autonomous robots] is increasing fast, and it seems unlikely that we will be able or willing to abstain from their use in the future" [1: 176]. In this rush to build robots made in our image, we will only get halfway, and thus end with machines that make their own decisions yet

experience neither pleasure nor pain. The combination that motivates the philosophical literature is thus essentially this: autonomous + unfeeling = responsibility gap. While Matthias did not feel any need to stress the point at the time, Königs' compromise now requires us to go one step further and openly state an analytic truth that everybody else grasped implicitly: responsibility + gap = *problem*.

I recall being patently indifferent when I was told that there are eight planets in our solar system as opposed to nine. My indifference was fully compatible with my grasp of the concepts "eight," "nine," and "planet." Such indifference could not apply, however, if I were told that there is one more sociopath in my neighborhood. It is not just a personal dislike of sociopaths that is involved. The issue is semantic-*cum*-logical: if one grasps what the concept describes and one grants that something satisfies the description, it *has* to be a problem. We have normative vocabulary precisely so that affirming normative propositions has practical weight. Hence, to countenance gaps in responsibility (at a purely descriptive level) is to confess that one would rather not confront them (at a normative level).

Königs, however, wants to accept the left-hand side of the responsibility + gap = problem formula and reject its right-hand side. One could perhaps do this by redefining what those terms mean. Yet, to his credit, Königs deploys the concepts as ethicists and lay people standardly understand them. He merely resists the evaluation that everybody else takes those concepts to entail. My hope is that, whatever else we may disagree on, we can agree that this admixture of commitments is untenable.

4. What "gap" (metaphorically) means

The word "gap" typically indicates a spatial emptiness or void, so when used in an ethical context, it must be a stylistic flourish. In the context that interests us, it signals a negation or privation of some sort. Hence, to say that there is a "responsibility gap" just is to say that there is no responsibility or responsible party.

Words can be combined in more ways than concepts can, but it is the underlying concepts, not the words, which determine whether a stance is feasible. My central point can thus be arrived at from another angle.

We can all agree that, had Matthias fancied different words, he could have just as easily called it "the responsibility *problem*." Indeed, the sentence immediately following his introduction of the expression "responsibility gap" describes it as a "problem" [1: 177]. Had Matthias opted for the label "responsibility *problem*," Königs would not be able to acknowledge its existence while denying that it is a problem. The contradiction at hand would be too overt. I fail to see how Matthias' superficial word choice generates the conceptual possibility that Königs wishes to occupy (in this regard, Matthias' coining may be a victim of its own viral success, since you never want your label to be so catchy that it obscures the substance of what is involved).

5. How problematic must a "problem" be?

Reconstructing Königs' reasons, it may be that he deems the responsibility gap unproblematic because he introduces stipulations that make the situation rarer and more manageable. As he sees it, "[i]f negligent, reckless or malicious behavior led to an autonomous system causing harm, whoever engaged in this behavior—the

manufacturer, the programmer, the operator, etc.—clearly is blameworthy" [11: 6]. Superficially, this seems right. Yet, when we examine the situation more carefully, we find that this stance deploys the concept of "autonomy" without fully considering how it changes things.

Königs mentions a "highly autonomous vehicle" and says "[t]he fact that the vehicle acts autonomously and unpredictably does not mean that the engineer is not responsible" [11: 4]. He thus wants engineers to do "more to foresee and limit the risk of harm" [11: 4]. This harm-prevention intent is laudable, but how could anyone possibly meet Königs' request, given the autonomy at hand? By definition, one cannot "foresee and limit" what autonomous agents do (the relationship that police and parents have with adult children would be very different if it were otherwise). We thus need to be clear about the type of robot considered. Is the vehicle discussed by Königs "highly" autonomous or *truly* autonomous? If a manufacturer can be proven to have had some input in a robot's particular deed, then that robot was not truly autonomous. We remain free to use the word "autonomy" in a loose manner, but it is only autonomy in its strict free-will-like version that generates responsibility gaps.

When a robot is autonomous, its specific actions are causally uncoupled from that robot's point of origin. Humans witnessing autonomous robotic decisions are not connected to those decisions in a way that satisfies what Taddeo and Blanchard call the "causality condition," which says that in order for an agent to deserve blame "there has to be a causal connection between the decision/action of the agent and their effects" [14: 4]. Even a company run by saints would have no say in what an autonomous robot does, once it is released in the wild. To dramatize this point: if an autonomous robot spontaneously murdered innocent people, it would make no difference whether it was built by the Boston Dynamics, the Taliban, or Disney. So, while I agree with Deborah Johnson that "[t]here are good reasons for staying with human responsibility," it cannot solve all our problems to put "pressure on developers to ensure the safety and reliability of such devices" [15: 714] when the device in question is autonomous. We thus need a better fix than blaming the nearest bad actor.

Königs stipulates that "[r]esponsibility gaps can plausibly arise only in situations in which the negative outcome is not *due to* carelessness or malice" [11: 4; emphasis added], but this stipulation fails to consider that "due to" talk is inapplicable to agents capable of making their own moral laws (from the Greek *auto-nomos*). Once we reach a certain level of sophistication the folks in charge are no longer in charge. What we *can* do, however, is refrain from releasing autonomous robots in the first place or demand moral answerability of the humans who release them. Importantly, whatever pressure we exert must be applied *beforehand* since, once released, programmers and developers can justifiably exonerate themselves by citing the lack of any causal link. Companies will thus prioritize safety or severely limit AI use only if they publicly accept real consequences for whatever negative outcomes ensue [16].

We humans can build machines whose "artificial intelligence" leaves us no say in their particular decisions. As puzzled as we may be by this problem of our own making, it remains a problem. I have been emphasizing how Königs does not regard it as problematic, but what generates an untenable tension is his recognition that the responsibility gap exists. Surely we can grant that such a gap, however rare, leaves our ordinary blaming and holding-responsible practices unfulfilled or in need of revision. We might be able, as Champagne and Tonkens [4], Kiener [5], Gunkel [7], Gogoshin [8], and Søvik [9] suggest (from different angles), to reshape and redirect those blaming and holding-responsible practices to meet a surprising new issue. Fashions to

the side, humans do not rework their mores and norms unless they are given a real cause to do so. Hence, the mere fact that we now feel a need to think or act differently as a result of autonomous robots attests to the gap being problematic to some degree.

We have to redirect our blame or at least justify why we don't redirect it. Königs may not lose sleep over the redirection option, but justifying why the moral practices already in place suffice did merit a considerable slice of his philosophical attention. Hence, at minimum, the responsibility gap qualifies as a problem in the same way that, say, the mind-body problem in philosophy of mind is a problem.

6. No problem? No tangible recommendation either

Unlike the mind-body problem, problems about responsibility have real-life urgency. When an autonomous robot misbehaves, people get hurt. Königs remarks that "[t]he intuitively most compelling concern about responsibility gaps revolves around their possible harmful consequences" [11: 5]. Robots are human-made artefacts, so unlike natural events like hurricanes, someone needs to be held responsible for the hurt generated. Usually, "[w]hen we judge a person to be responsible for an action," one of the things that we mean is that "the person […] is rightly subject to a range of specific reactive attitudes like resentment, gratitude, censure, or praise" [1: 175]. The question is how to justly pinpoint the culpable party. What makes this question tricky is that the robot which caused the calamities exhibits autonomy. Matthias was the first to remark that, as a result of this autonomy, "the manufacturer/operator of the machine is *in principle* not capable of predicting the future machine behaviour" [1: 175].

If holding humans responsible has good consequences (like deterrence) and robototic autonomy prevents us from holding humans responsible, then robotic autonomy robs us of those good consequences. Königs thus grasps the common worry that "if nobody is responsible, people are under no pressure to minimize harm and damage," which in turn "erode[s] a socially beneficial incentive structure" [11: 5].

Now, according to Königs, "that an autonomous system's high degree of autonomy exculpates" negligent, reckless or malicious behavior on the part of a manufacturer, programmer, or operator "is precisely what is so implausible to assume" [11: 6]. I find it inappropriate to label this an "assumption" (let alone an implausible one), since it received an extensive defense. Matthias' original discussion of artificially intelligent neural nets went to great lengths to demonstrate how and why it is impossible to link the decisions of an autonomous robot to the decisions of a human [1]. In his influential piece, Matthias gives the (real, not made-up) example of an adaptive elevator system which uses AI to minimize how long users wait before getting a car. Although "it might be possible to prove mathematically that eventually the used algorithm will converge to some optimal behavior," the AI component means that it is "no longer possible for the manufacturer to predict or control the specific behaviour of the elevator in a given situation" [1: 176].

It is beyond dispute that robots endowed with AI behave in ways that allow us to discern attractors. The AI elevator will *tend* to yield short wait times, just as the AI soldier will *tend* to pick correct targets. However, to paraphrase what Leibniz said with regards to free will, such attractors incline without necessitating. The observable patterns of behavior exhibited by an autonomous robot are the result of training instead of direct programming, so this machine learning leaves room for spontaneous actions completely at odds with what we expect or intend the robot to do. One recent discovery

that has been counter-intuitive is that larger neural nets aggravate, not alleviate, this possibility of unpredictable misalignments with human values and expectations [17: 2]. So, if we value control, we should be making our machines dumber, not smarter.

When a robot is explicitly built to be artificially intelligent, its independent choices are not a bug but a feature. It should not be tendentious, then, to admit that we always risk being surprised by what the robot does. This loss of control changes everything. What evidence do we normally have for calling a robot manufacturer negligent, reckless or malicious? A robot's harm-causing malfunction, of course. But, if no link can be established between that malfunction and what the manufacturer did or intended, those assessments of blameworthiness and pejorative labels become misplaced.

Adopting a stance similar to Königs' compromise, Huzeyfe Demirtas seeks to deflate ethicists' worries with the following analogy: "Suppose instead of deploying AI, Soldier shoots an arrow into the warzone and Victim dies as a result. If Soldier took all the reasonable precautions, she's not responsible for Victim's death" [18: 120]. This analogy begs the question. Demirtas talks of an AI that "unexpectedly malfunctions" [18: 121], but the notion of autonomy means that we must wrap our minds around the more challenging (and arguably unprecedented) idea that an AI can *unexpectedly function*. So, if we truly wish to calibrate our intuitions, an analogy more apt than flying arrows would be the cartoon revolver from the movie *Who Framed Roger Rabbit*, where the six-chamber cylinder is loaded with little cowboy characters. Would we equip police officers with those sorts of guns? And, if a bullet-person spontaneously decided to head in a direction of its own choosing—as happens in the movie—would we be justified in holding the officer responsible? Upon witnessing the cowboy bullets take matters into their own hands, the movie's protagonist Eddie Valiant tosses the revolver to the curb. When it comes to AI, an argument may be made that we should do the same, especially since smart but non-autonomous robots will satisfy most needs.

Now, if Königs were to oppose the strong idea that these gaps in responsibility pose an *insoluble* problem, I would understand that (and oppose it too). This is not, however, what he says. He claims that the responsibility gap is not a problem *tout court*. What would someone who adopted this view think about, say, autonomous robots that go rogue in war? Demirtas contends that a drone falsely identifying a non-combatant as a threat and killing him "is no different than any other unfortunate case where things just go wrong despite all precautionary measures. Though even if one disagrees, this isn't troubling for me" [18: 116]. No policy recommendation could be extracted from this sanguine indifference. Königs says that "[t]he legal framework should piggyback on what is just" [11: 6]. I agree. But, if no special injustice occurred, then no special ethical or legal tools are required. One may thus regard the responsibility gap as "not morally problematic in a way that counts against developing or using AI" [18: 10].

Right now, some open letter signatories to the side, few think that AI is morally problematic in a way that counts against developing or using it. Although I will stay silent on whether continuing this status quo constitutes an improvement or setback, the fact that Königs' compromise blends well with present trends is worth bearing in mind.

7. Conclusion

The foregoing demonstration that Königs' compromise is untenable employed an analysis of what constituent concepts mean. Just as "unmarried" plus "male" get you "bachelor" for free, "responsibility" plus "gap" get you "problem" for free. It violates

the meanings involved, then, to acknowledge that responsibility is lacking yet claim indifference.

Redundancy tends to threaten debate spaces that become crowded, so as was said at the outset, participants would celebrate the addition of a new possibility. Although provocative suggestions can often be rejuvenative, they sometimes result in stances that cannot stand—just as developers venturing atop land reclamation risk building on unstable foundations. The contradiction involved in Königs' compromise may not be as overt as a head-on P and not-P clash, but it can nevertheless be revealed by attending more carefully to the concepts employed.

Königs is aware that his yes-to-existence and no-to-problem compromise is vulnerable on that front. He observes that, were one to stress the analytic entailments I have stressed, "the problematic nature of responsibility gaps would be built into the definition. For there to be a responsibility gap would mean for there to be a *problematic* absence of responsibility" [11: 3; emphasis in original]. In that case, he says, his response "would quite simply be that responsibility gaps do not exist at all" [11: 3].

Whatever its ultimate merit, such a move would at least betoken a return to solid ground. So, if my arguments can compel Königs to switch from his compromise to an outright denial of the responsibility gap, I will consider that incremental progress.

References

[1] Matthias A. The responsibility gap: Ascribing responsibility for the actions of learning automata. Ethics and Information Technology, 2004; 6(3): 175-83. doi:10.1007/s10676-004-3422-1

[2] Sparrow R. Killer robots. Journal of Applied Philosophy, 2007; 24(1): 62-77. doi:10.1111/j.1468-5930.2007.00346.x

[3] Arkin RC. The case for ethical autonomy in unmanned systems. Journal of Military Ethics, 2010; 9(4): 332-41. doi:10.1080/15027570.2010.536402

[4] Champagne M, Tonkens R. Bridging the responsibility gap in automated warfare. Philosophy and Technology, 2015; 28(1): 125-37. doi:10.1007/s13347-013-0138-3

[5] Kiener M. Can we bridge AI's responsibility gap at will? Ethical Theory and Moral Practice, 2022; 25(4): 575-93. doi:10.1007/s10677-022-10313-9

[6] Burri S. What is the moral problem with killer robots? In: Strawser BJ, Jenkins R, Robillard M, editors. Who should die?: The ethics of killing in war. Oxford University Press; 2018. p. 163-85. doi:10.1093/oso/9780190495657.001.0001

[7] Gunkel DJ. Mind the gap: Responsible robotics and the problem of responsibility. Ethics and Information Technology, 2020; 22(4): 307-20. doi:10.1007/s10676-017-9442-4

[8] Gogoshin DL. Robot responsibility and moral community. Frontiers in Robotics and AI, 2021; 8: 768092. doi:10.3389/frobt.2021.768092

[9] Søvik AO. How a non-conscious robot could be an agent with capacity for morally responsible behaviour. AI and Ethics, 2022; 2(4): 789-800. doi:10.1007/s43681-022-00140-0

[10] Tigard DW. There is no techno-responsibility gap. Philosophy and Technology, 2020; 34(3): 589-607. doi:10.1007/s13347-020-00414-7

[11] Königs P. Artificial intelligence and responsibility gaps: What is the problem? Ethics and Information Technology, 2022; 24(3): 36. doi:10.1007/s10676-022-09643-0

[12] Königs P. No wellbeing for robots (and hence no rights). American Philosophical Quarterly, in press.

[13] Königs P. What is techno-optimism? Philosophy and Technology, 2022b; 35(3): 63. doi:10.1007/s13347-022-00555-x

[14] Taddeo M, Blanchard A. Accepting moral responsibility for the actions of autonomous weapons systems—a moral gambit. Philosophy and Technology, 2022; 35(3): 1-24. doi:10.1007/s13347-022-00571-x

[15] Johnson DG. Technology with no human responsibility. Journal of Business Ethics, 2015; 127(4): 707-15. doi:10.1007/s10551-014-2180-1

[16] Champagne M, Tonkens R. A comparative defense of self-initiated prospective moral answerability for autonomous robot harm. Science and Engineering Ethics, 2023; 29(4): 27. doi:10.1007/s11948-023-00449-x
[17] Alexander P, Bhatia K, Steinhardt J. The effects of reward misspecification: Mapping and mitigating misaligned models. arXiv:2201.03544v2 [Preprint]. 2022 [cited 2024 May 30]. Available from https://arxiv.org/abs/2201.03544v2
[18] Demirtas H. Responsibility internalism and responsibility for AI [dissertation]. Syracuse (NY): Syracuse University; 2023.