Journal of Philosophical Theological Research
Print: 1735-9791
Online: 2538-2500

Normative Requisites of Forgiveness and AI Agency

Ishtiyaque Haji

Professor of Philosophy, Department of Philosophy, University of Calgary, Canada,
ihaji@ucalgary.ca

Abstract

Regarding forgiveness, there is disagreement on the following "contrast pairs." (i) The victim (the person who forgives) forgives the offender (the person who is forgiven) for doing something only if it is morally wrong for the offender to do this thing; or even if she has done no wrong, the victim forgives the offender only if she takes herself to have done wrong. (ii) The victim forgives the offender for doing something only if she is blameworthy for doing this thing; or even if not blameworthy, the victim forgives the offender only if she takes herself to be blameworthy for doing this thing. In this paper, first, I appeal to "Frankfurt examples" to argue that the belief options – the second members – of each contrast pair are more tenable. Second, I discuss why even these belief options threaten forgiveness if determinism is true in virtue of determinism's imperiling wrongness. Third, I broach the issue of whether artificial intelligence agents ("AI agents") can be apt candidates of forgiveness. I synopsize conditions AI agents must satisfy to be fitting subjects of forgiveness, emphasizing agency requisites for blameworthiness and wrongdoing. The summary's principal aim is to animate further inquiry into whether AI agents can be agents of the sort that forgiveness presupposes.

Original Research

Keywords:

artificial intelligence agency, blameworthiness, Frankfurt examples, forgiveness, wrongness.

Received: 2026/02/12; **Received in revised form:** 2026/04/16; **Accepted:** 2026/05/12; **Published online:** 2026/06/05

Haji, I. (2026). Normative Requisites of Forgiveness and AI Agency. *Journal of Philosophical Theological Research*, (Special Issue on AI, Consciousness, and Human Agency), 28(1), 19-40. https://doi.org/10.22091/jptr.2026.16330.3588

Copyright © The Author

Introduction

Assuming that there are or can be artificial intelligence agents ("AI agents") who are minimal agents – entities capable of performing intentional mental or bodily actions – can AI agents be apt candidates of forgiveness? Answering this question requires multifaceted inquiry into, among other things, whether AI agents can do moral wrong, have beliefs about wrongdoing, be blameworthy for some of their actions, or have beliefs about whether they are blameworthy for their actions. This is because, reasonably, there are conceptual connections among forgiveness, wrongness or belief in wrongness, and blameworthiness or belief in blameworthiness. In this paper I first explore these putative connections. I then briefly summarize a non-exhaustive list of the preconditions AI agents must satisfy to be appropriate candidates of forgiveness, emphasizing agency requisites for blameworthiness and wrongdoing. The primary aim of this synopsis is to motivate further inquiry into whether, in addition to being agents in the minimal sense, AI agents can be the type of augmented agent that forgiveness presupposes.

Forgiveness and Dual Contrast Pairs

Consider these "contrast pairs" regarding putative requirements of forgiveness:

Wrongness Contrast Pair

(i) Objective Member: The victim (the person who forgives) forgives the offender (the person who is forgiven) for doing something only if it is morally wrong for the offender to do this thing. This requirement may be coupled with the additional one that the victim believes that the offender has done wrong.[1]

(ii) Subjective Member: The victim forgives the offender for doing something only if the offender believes that it is morally wrong for her to do this thing and the victim believes that the offender takes herself to have done wrong.

The subjective members may be qualified in various ways, for example, regarding (i), one may venture that the offender justifiably believes that she has done wrong. Such qualifications play no substantial role in what is to come

1. As Margaret Holmgren explains, "Where there is no wrongdoing, there is nothing for the injured person to forgive, although there may be much for her to accept….Forgiveness is an issue only when a moral agent commits an offense, without justification, and in the absence of an exculpating excuse" (2012, 35). (See, e.g., also Murphy & Hampton, 1988, 20; Holmgren, 1993; Hughes, 1993; Hieronymi, 2001, 546; Nelkin, 2013, 175; Allais, 2008; 2014, 41).

because any such qualified view entails the unadorned belief view.

Blameworthiness Contrast Pair

(i) Objective Member: The victim forgives the offender for doing something only if the offender is morally blameworthy for this thing. If true, it may be proposed that this precondition goes together with the provision that the victim believes that the offender is blameworthy.[1]

(ii) Subjective Member: The victim forgives the offender for doing something only if the offender believes that she is blameworthy for this thing, and the victim believes that the offender takes herself to be blameworthy for this thing.

Again, (ii) may be restricted in certain ways. For instance, it may be advocated that the victim nonculpably believe that the offender is blameworthy. Such qualifications will be largely ignored because they are inconsequential to the issues to be discussed below.

These pairs may compete with others. With the first, one may submit that the offender can be forgiven for something she did that is bad but not impermissible.[2] As for the objective member of the second pair, determinism is the thesis that at any instant exactly one possible future is compatible with the state of the universe at that instant and the laws of nature (van Inwagen, 1983, p. 3). One may be tempted to defend the view that even if determinism undermines blameworthiness, determinism does not preclude forgiveness.[3] In this paper, I disregard these other options. My primary focus is to ascertain which member of each "contrast pair" has the upper hand. The winner of each pair will have to contend with rivals. In what follows, first, I appeal to Frankfurt examples to support the subjective variants of each pair. Then I explain why there are reasons to believe that determinism threatens these variants largely owing to determinism's, at least prima facie, imperiling moral

1. Dana Nelkin contends that when "one forgives an offense, one is committed to the proposition that the offender acted freely and was responsible for the offense" (Nelkin, 2013, p. 183). Similarly, Brandon Warmke and Michael McKenna emphasize that "directed forgiveness presumes that the person forgiven is indeed blameworthy, and that the victim had blamed the person previously, or at least held her morally responsible and blameworthy" (Warmke & McKenna, 2013, 205). Warmke also remarks that forgiveness "does not change the fact that one did something for which one is guilty and is liable to punishment by the appropriate moral, civil or even divine authorities" (2016, 13).

2. I use "wrong" and "impermissible" interchangeably.

3. See, e.g., Pereboom, 2001; 2014. Pereboom has recently defended the view that forgiveness presupposes blameworthiness ("protest blameworthiness") of a consequentialist variety (2021, pp. 103-104).

wrongness. Finally, I revert to the issue of whether AI agents can be forgiven.

Forgiveness and Impermissibility

Frankfurt examples provide support for the wrongness contrast pair's subjective member that forgiveness requires belief in wrongdoing and not wrongdoing *per se*. Typically, these examples have been invoked to overturn the principle of alternate possibilities (*PAP-R*): agents are morally responsible for what they have done only if they could have done otherwise. These examples may be divided into two stages. In stage 1 of case *Forgiveness,* Laleh ridicules Arman for questionable reasons. In this stage, she could have refrained from deriding Arman, and it is impermissible for her to deride him, and she is blameworthy for doing so. In stage 2, a "rerun" of stage 1, a counterfactual intervener – Omid – who fails to show his hand, would have forced Laleh to ridicule Arman had she displayed an involuntary sign that she was about to refrain. But Laleh acts no differently in this stage than in the former. It looks like she is blameworthy for her relevant behavior in stage 2 even though she could not have done otherwise. Hence, Frankfurt partisans conclude, *PAP-R* is false (see Frankfurt, 1969).

Add to *Forgiveness* that Arman gets wind of Laleh's untoward behavior (in stage 2), and he forgives her. Without good reason to believe otherwise, it appears that there is nothing suspect about this instance of forgiveness.[1] If you agree, you have reason to discard the objective view that to be forgiven for something, it is impermissible for the offender to do it. Why so? Although Arman, as assumed, forgives Laleh for ridiculing him, in stage 2 it is not impermissible for Laleh to ridicule him because no one can do wrong unless one could have done otherwise, and in stage 2 Laleh could not have refrained from deriding Arman.

The view that impermissibility requires freedom to do otherwise rests on two credible principles:

> *Ought Implies Can* (*OIC*): If it is obligatory for you to do *A*, then you can do *A* (and if it is obligatory for you not to do *A*, then you can refrain from doing *A*).[2]

1. One may propose that, here, one does forgive but the forgiveness is not appropriate or fitting. But then the relevant issue concerning wrongness and forgiveness should be taken to address appropriate forgiveness.

2. E.g., Feldman, 1986; Zimmerman, 1996; Haji, 2016; Vranas, 2007; Littlejohn, 2012. I omit temporal indices here. I take the agent- and time-relativized canonical form of obligation

Just as you cannot be responsible for something unless you are free with respect to it and, thus, you can do it, nothing can be obligatory for you unless you are free regarding it and, so, you are able to do it.

> *Equivalence*: It is obligatory for you to refrain from doing *A* if, and only if, it is impermissible for you to do *A*.

In the estimation of many, this "ought not" is equivalent to "impermissible" principle has the status of a morally deontic axiom or theorem.[1]

> *OIC* and *Equivalence* jointly entail:
>
> *Impermissibility-Implies-Can-Avoid* (*IAvoid*): If it is impermissible for you to do *A*, then you can avoid doing *A*.

It is not impermissible for Laleh to ridicule Arman since she cannot avoid ridiculing him in stage 2. However, Arman forgives her for her behavior. Hence, it is not a precondition of forgiveness that the offender has done wrong. If this precondition is false, it appears that there is little reason to accept the corollary that forgiveness presupposes that the victim believes that the offender has done wrong. In short, the wrongness contrast pair's objective member is false.

A caveat is in order. Suppose you have initial (or "pretheoretical") inclinations regarding which member of each contrast pair under scrutiny is true. For example, it might be that you initially judge that forgiveness surely requires that the offender does wrong. Then you may be drawn to claiming that in Frankfurt examples the victim does not forgive because the offender's action – in *Forgiveness*, Laleh's ridiculing Arman – is not impermissible for her. A relevant comparison is that you may antecedently judge that *PAP-Blame* – the principle that blameworthiness requires freedom to do otherwise – is true. Then Frankfurt examples won't move you to renounce *PAP-Blame*; if you are already committed to *PAP-Blame*, you should insist that in *Forgiveness* Laleh is not blameworthy for ridiculing Arman in stage 2. Similarly, if you antecedently judge that forgiveness presupposes blameworthiness, and blameworthiness requires freedom to do otherwise, Frankfurt examples won't convince you that the victim forgives the offender in Frankfurt cases. However, this would not be the proper way to proceed if the very inquiry concerns which member of each contrast pair is true. On pain

---

statements to be this: at *t*, it is obligatory for *S* to do *A* at *t** (where *t** may be *t* or later than *t*).

1. E.g., Belzer, 1998; Hilpinen & McNamara, 2013, 43; McNamara, 2019; Rönnedal, 2009, 28-29. On *Equivalence*, see, also, Pereboom, 2001, 146-147; and Haji, 2002, 52-53.

of begging the question in one direction or other, it should not be these antecedent judgments that decisively inform your pertinent verdicts regarding forgiveness and impermissibility or forgiveness and blameworthiness in Frankfurt examples.

I now address two objections to this appeal to Frankfurt examples against the wrongness contrast pair's objective member. The first is that its not being impermissible for Laleh to deride Arman in stage 2 yields the absurd result that it is permissible for her to engage in this behavior. However, the inference from "not impermissible" to "permissible" is unwarranted. To see this, imagine that in an amended Frankfurt-like example, Fatemeh is not blameworthy for unjustifiably rebuffing Azad in stage 2 because Omid forces Fatemeh's hand; Fatemeh fails to exercise *responsibility-level control* when she snubs Azad. It would be a mistake to conclude that since Fatemeh is not blameworthy, she is either praiseworthy or "neutrally" responsible for her rebuffing in stage 2. Without the control that responsibility demands, she would not be responsible at all. Analogously, should Laleh fail to exercise *obligation-level control* in ridiculing Arman, it is neither obligatory nor permissible for her to ridicule Arman, in addition to its not being impermissible for her to ridicule him. Rather, Laleh's act of ridiculing Arman is amoral for her: ridiculing Arman is not permissible, impermissible, or obligatory for her.

My position here rests on the principle, *Obligation/Freedom*, that obligation requires control. This principle is homologous to the well accepted principle, *Responsibility/Freedom*, that responsibility requires control. You may intentionally perform an action but in so doing, lack obligation- or prohibition-relevant control in performing it. This action which would otherwise be obligatory (or impermissible) for you is neither. So, what may be said to support *Obligation/Freedom* and, hence, to reject the view that actions are either permissible or impermissible for any moral agent? What may be offered to bolster the view that there is such a category as the *amoral* for moral agents? Although I accept *Responsibility/Freedom* I have no direct argument for it (that is, I have no argument with this principle as its conclusion), nor have I seen any plausible *direct* argument for it. I endorse *Responsibility/Freedom* because it has historical backing, it is intuitively plausible, and reflection on a range of cases persuades me that responsibility requires freedom. I accede to *Obligation/Freedom* for similar reasons: it has historical precedent, it is intuitively credible, attention to relevant cases convinces me that it is reasonable, and accepting it, along with other apt propositions, enables me to achieve reflective equilibrium regarding an overall account of obligation (see Haji, 2019). Here is one sort of case I find especially revealing. Grant that Laleh could not have done otherwise in *Forgiveness*.

Then it isn't impermissible for her to ridicule Arman in stage 2 (given *OIC* and *Equivalence*). I've proposed that it's not impermissible for her to ridicule Arman because she could not have done otherwise.

Suppose one claims that since it is not impermissible for her to ridicule Arman, it must be permissible for her because an act is either permissible or impermissible for any moral agent. First, this result is highly counterintuitive. (To drive home this contention, imagine that the action Laleh cannot avoid performing in stage 2 is pulling the trigger (thereby ensuring that an innocent person is killed)). Second, if you think this action is permissible, you are committed to this asymmetry: Whereas impermissibility requires freedom (freedom to do otherwise), permissibility does not. But I can't think of a plausible reason for the asymmetry in the control requirements of impermissibility and permissibility. Nor should we expect this sort of asymmetry, given the intimate conceptual connection between obligation and impermissibility that *Equivalence* exposes. Elaborating, *A* is optional for you if, and only if, both *A* and not-*A* (refraining from doing *A*) are permissible for you. If it is permissible for you to do *A*, then either it is obligatory for you to do *A*, or it is optional for you to do *A*. Assume some act, *B*, is impermissible for you. Then, you can do otherwise than *B* (since impermissibility requires avoidability), and you have a negative obligation concerning *B* – you ought not to do *B* – since "ought not" is equivalent to "impermissible." Barring strong reason to think otherwise, the freedom requirements of an obligation to do something (a positive obligation) and an obligation to refrain from doing something (a negative obligation or a prohibition) are symmetric. Now suppose, first, *A* is permissible for you because *A* is obligatory for you. Then, again, in the absence of reasons to believe otherwise, why think that in this sort of case, the freedom requirements of permissibility differ from those of obligation, and so differ from those of prohibition or impermissibility on the supposition that negative and positive obligations have symmetric freedom requirements? Second, if in this sort of case, the freedom requirements of permissibility don't diverge from those of impermissibility, why believe that we should expect a divergence in these requirements in the remaining cases in which *A* is permissible for you because *A* is optional for you? The foregoing are not decisive reasons for accepting *Obligation/Freedom* and the proposal that the freedom requirements of permissibility, impermissibility, and obligation are symmetric, but they are not without merit either.

I next turn to a second objection to the view that in *Forgiveness* it's not impermissible for Laleh to ridicule Arman in stage 2. The general worry is that there are different kinds of ability, and if different kinds of ability track different kinds of control, couldn't Frankfurt cases be a reason to *reject* a

particular interpretation of *OIC* that appeals to a sort of ability (specific ability to do otherwise) that Laleh ostensibly lacks in stage 2, but not to reject the kind of control that blameworthiness requires? In response, distinguish between two broad classes of ability, specific and general ability. To illustrate the difference, Shirin, on the road to the market, cannot steal any dates because along the way there are no dates to steal. Although she lacks the specific ability to steal, she has the general ability; even though she has no opportunity to steal, she retains her thieving talents and skills. Entering the market, Shirin plies her trade: she pilfers some dates, and in so doing exercises her specific ability to steal.[1] Informally, and appealing to a different example, you have the general ability to ride a bike insofar as you have the intrinsic powers – balance, motor control, skills and so forth – to ride. Your ability is increasingly specific to the extent that you also have the agreeable state – the physiological and psychological capacities – to ride (see Clarke, 2015; Vihvelin, 2004). Your ability is more specific still to the extent that you additionally have the opportunity – access to a bike and window for use – to ride. Roughly, you have the specific ability to ride a bike if, and only if, you have the intrinsic powers, are in the agreeable state, and have opportunity to ride.

Revisiting *Forgiveness*, with Omid in the wings, and excluding other problems with Frankfurt examples, independently of confirmation by an analysis of specific ability, it's plausible that although Laleh has the general ability to refrain from ridiculing Arman in Stage 2, she lacks the specific ability to refrain.[2] Regarding the objection under scrutiny – that the control requirements of blameworthiness and obligation (or impermissibility) track different sorts of ability – maybe the objector will propose, first, that Frankfurt cases like *Forgiveness* reveal that in stage 2 Laleh lacks the specific ability to refrain from ridiculing Arman but not the specific ability to ridicule Arman. In addition, the objector may insist that blameworthiness for an action requires that one have the specific ability to perform the action but not the specific ability to refrain as Frankfurt examples show. In contrast, obligation requires general and not specific ability or at least some kind of ability that is much more akin to general ability and much less like specific ability. Laleh retains the general ability both to ridicule and to refrain from ridiculing Arman even in stage 2 with the counterfactual intervener on duty. Hence, there is no bar to maintaining that in stage 2 it *is* impermissible for Laleh to ridicule Arman.

---

1. For more on the distinction between specific and general ability, see, e.g., Haji, 2016, p. 213; Hebert 2016.

2. One may propose that even in stage 2, Jones has the *specific* ability to refrain from stealing. See, e.g., Nelkin 2011, 66-71; and Vihvelin 2013, 195-196. For responses to Nelkin and Vihvelin, see Haji 2019, 190-200.

In reply, if there is a requirement of alternate possibilities for obligation and impermissibility, it is implausible that the requisite ability to do otherwise is general and not specific. Proponents of *OIC* cannot sensibly be taken to propose that an obligation to do something presupposes only the general and not the specific ability to do it. Imagine you are on lifeguard duty and a child in the ocean is in serious trouble. You cannot save the child – you don't have the specific ability to save the child – because sharks in the water between you and the child would eat you if you attempted to save the child. Although you have the general ability to save the child – you have the skills and know-how to save – you lack the opportunity to save. Hence, it would be unreasonable to insist that you morally ought to save the child even though you cannot. If you ought to do something, you can do it, and "can" here expresses specific ability. Excluding special reason to believe otherwise, if "ought" and "impermissibility" entail "can avoid," the "can" in "can avoid" also expresses specific ability. In Frankfurt examples, however, agents like Laleh in stage 2 lack the specific ability to refrain from doing what they cannot avoid doing. In addition, should the objector think that Frankfurt cases show that obligation does not require specific ability but some other sort of ability, expose what this other variety of ability is – surely, it can't be general ability – and then I will be happy to attempt to assess what is on offer.

To summarize, the offender's relevant action in pertinent Frankfurt cases (for example, Laleh's ridiculing Arman in *Forgiveness*) is not impermissible for her (given *OIC* and *Equivalence*). You might, next, reason in these steps:

Argument 1

(1W) Since neither offender nor victim act any differently in a Frankfurt case than they do in a non-Frankfurt version of the case with no fail-safe mechanism (essentially stage 1), if forgiveness is not problematic in the non-Frankfurt case, it should not be problematic in the Frankfurt case either. (2W) As the victim still legitimately forgives in stage 2, forgiveness does not presuppose impermissibility. (3W) Given the choice between the objective impermissibility requirement and the subjective belief-in-impermissibility requirement, the latter is to be preferred.

When we forgive, however, the view that the offender has done something that is negatively valenced is not withdrawn. Something of a negative nature remains constant. If it is not wrongdoing that persists, one may recommend instead that what remains constant is the victim's belief that the offender has done wrong and that the offender is either blameworthy or the victim continues to believe that the offender is blameworthy.

Forgiveness and Blameworthiness

Turning now to the blameworthiness contrast pair, which of its members is more plausible: necessarily, to be forgiven for something (i) the offender is blameworthy for this thing and the victim believes that the offender is blameworthy for it, or (ii) the offender takes herself to be blameworthy for this thing and the victim believes that the offender takes herself to be blameworthy? An argument against the latter, which parallels the argument in the previous paragraph against the objective wrongness view, is problematic. To explain, appealing to *Forgiveness*, ponder the following against what may be labelled the "objective blameworthiness view":

Argument 2

(1B1) Since neither offender nor victim act any differently in a Frankfurt case than they do in a non-Frankfurt version of the case, if forgiveness is not problematic in the non-Frankfurt case (recall, in stage 1, the agent *is* blameworthy), it should not be problematic in the Frankfurt case either. (2B1) As the victim still legitimately forgives in stage 2, forgiveness does not presuppose blameworthiness. (3B1) Given the choice between the objective blameworthiness requirement and the subjective belief-in-blameworthiness requirement, the latter is to be preferred.

Frankfurt partisans should reject (2B1), insisting that Laleh remains blameworthy in stage 2 even though she could not have refrained from ridiculing Arman. After all, Frankfurt examples have been commissioned to undermine, among other versions of *PAP-R*, the principle (*PAP-Blame*) that you are blameworthy for doing something only if you could have done otherwise. How do we proceed in the face of this difficulty?

Let's look more closely at an alleged conceptual connection between blameworthiness and impermissibility, the principle that you cannot be blameworthy for something unless it is impermissible for you:

> *Blameworthiness Requires Impermissibility* (*BRI*): You are blameworthy for *A*-ing only if it is impermissible for you to do *A*.[1]

Should Frankfurt examples like *Forgiveness* at least initially sway you against the view that blameworthiness requires avoidability, you have reason to reject *BRI*. Recall, from *OIC* ("ought" implies "can") and *Equivalence* ("ought not"

1. E.g., Smith, 1991, 271; Widerker, 1991, 223; Fields, 1994; Copp, 1997; Copp, 2003, 286-287; Fischer, 2006, 218; Arpaly, 2006, 91n3; Campbell, 2011, 33-34; Nelkin, 2011, 100, 104; Franklin, 2018, 35; Brink, 2021, 117.

is equivalent to "wrong"), we infer *IAvoid* (if it is impermissible for you to do *A*, then you can avoid doing *A*). In addition, *OIC*, *Equivalence*, and *BRI* jointly entail *PAP-Blame*: If you are blameworthy for doing *A*, it is impermissible for you to do *A* (*BRI*). If it is impermissible for you to do *A*, you can refrain from doing *A* (*IAvoid*). Hence, if you are blameworthy for doing *A*, you can refrain from doing *A* (*PAP-Blame*).[1] Provided you are convinced that Frankfurt examples do, indeed, undermine *PAP-Blame*, since *PAP-Blame* follows from the conjunction of *OIC*, *Equivalence*, and *BRI*, you should conclude that one or more of these principles is false. If you accept *OIC* and *Equivalence*, you should reject *BRI*.[2]

There are additional putative reasons to question *BRI*. Some have proposed that *BRI* falls to counterexample (See, e.g., Zimmerman, 1997). In addition, suberogatory acts are, roughly, morally permissible acts that are especially disvaluable in some way and for the doing of which you are blameworthy. One may attempt to undermine *BRI* by defending the possibility that some actions are suberogatory for agents.[3] Yet others may reject *BRI* by defending an account of blameworthiness that invalidates *BRI*.

Several candidates have been proposed as replacements for *BRI*.[4] I assume that the strongest of these ties blameworthiness to belief in impermissibility:

> *Blameworthiness Requires Belief In Impermissibility* (*BRBI*): You are blameworthy for *A*-ing only if you believe that it is impermissible for you to do *A*.[5]

1. E.g., Widerker, 1991; Haji, 1994; Nelkin, 2011, 100-101.
2. Fischer (2006, 217-222) proposes that Frankfurt examples impugn *OIC*. (see Haji, 2019, 39-42 for a reply.) If you reject *OIC*, the preceding and subsequent reasoning that appeals to it will not convince you.
3. Indeed, I think that you can be blameworthy for doing *A* even if *A* is not suberogatory for you but you believe that *A* is suberogatory for you. On the suberogatory, see, e.g., Mellema 1991; Driver, 1992; McNamara, 1996; Haji, 2016, pp. 88-91.
4. Here are some: (1) You are blameworthy for *A*-ing only if you have a sufficient reason not to *A* and it is reasonable to expect you to know this. See, e.g., Fischer, 2006, 221. (2) You are blameworthy for *A*-ing only if it is bad for you to do *A*, or, the state of affairs, *your doing A*, is bad (or, maybe, *A* is bad). (A belief variant of (2) says: You are blameworthy for *A*-ing only if you believe that the state of affairs, *your doing A*, is bad or you believe that *A* is bad, or you do *A* in light of the belief that *A* is bad.) (3) You are blameworthy for *A*-ing only if you do *A* for morally bad or questionable reasons. E.g., Hanser, 2005, 454, 456; Nelkin, 2011, 105; Pereboom, 2014, 143. (A belief variant says: You are blameworthy for *A*-ing only if you do *A* for what you take to be morally bad or questionable reasons.) See Haji, 2023, 31-37, for critical discussion of these principles.
5. E.g., Moore 1912, 101; Parfit 1984, 25; Thomson 1991, 295; Haji, 2016, 165-190; Zimmerman, 1997; Hanser, 2005, 454, 456. Pereboom rejects *BRI* (2014, 142).

For simplicity, I ignore variations of *BRBI* that qualify it in certain ways, for example, that your belief that it is impermissible for you to do *A* is nonculpable or justified because these variations don't bear on the subsequent discussion.

Remaining neutral on whether blameworthiness requires impermissibility (*BRI*) or mere belief in impermissibility (*BRBI*), I advance the following "bridge principle" that associates blameworthiness with impermissibility.

> *Blameworthiness/Impermissibility*: You are blameworthy for *A*-ing only if either it is impermissible for you to do *A* (*BRI*) or you believe that it is impermissible for you to do *A* (*BRBI*).[1]

Returning to whether forgiveness presupposes blameworthiness (the objective blameworthiness view) or belief in blameworthiness (the subjective blameworthiness alternative), assume that one of *BRI* or *BRBI* is true. Suppose *BRI* is true. Then one might advance this argument against the objective blameworthiness option:

Argument 3

(1B2) Since neither offender nor victim act any differently in a Frankfurt case than they do in a non-Frankfurt version of the case, if forgiveness is not problematic in the non-Frankfurt case, it should not be problematic in the Frankfurt case either. (2B2) As the victim still legitimately forgives in stage 2, forgiveness does not presuppose blameworthiness. Expanding on the rationale for (2B2), since the offender's act is *not* impermissible for her in stage 2 (as previously explained), and *BRI* (the principle that blameworthiness requires impermissibility) is assumed to be true, the offender is not blameworthy for what she does in stage 2. But if the victim still forgives in stage 2, forgiveness does not presuppose blameworthiness. (3B2) Given the choice between the objective blameworthiness requirement and the subjective belief-in-blameworthiness requirement, the latter wins.

Supposing *BRI* is true, it seems that (2B2) is no longer controversial provided that the offender's action is not impermissible for her in stage 2. However, there is still a left-over worry with *Argument 3*. Although *OIC* and *Equivalence* are relatively unproblematic, *BRI* is far more contentious. Indeed, as previously cautioned, Frankfurt cases themselves call *BRI* into question. (3B2) may, nevertheless, still be accepted because the alternative to *BRI* is

1. Maybe there is some complicated principle that conceptually ties blameworthiness to, for instance, all three of badness, bad reasons, and impermissibility. I assume any such principle (if there is one) entails one or both of *Blameworthiness/Impermissibility*'s disjuncts.

BRBI. If blameworthiness requires belief in impermissibility (*BRBI*) and not impermissibility *per se* (*BRI*), and if we grant that Arman does forgive Laleh in stage 2 and there is nothing suspect about this instance of forgiveness, then forgiveness is tied to belief in blameworthiness and not to blameworthiness *per se*. (Take it as given that Arman believes that Laleh is blameworthy for ridiculing him.)

Collecting results, grant that Arman *does* forgive Laleh in Frankfurt case *Forgiveness*. If *BRI* (blameworthiness requires impermissibility) is true, this should give us reason to accept the subjective belief-in-blameworthiness requirement on forgiveness as *Argument 3* shows. If *BRI* is false, and its alternative *BRBI* (blameworthiness requires belief in impermissibility) is true, this again gives us reason to accept the belief-in-blameworthiness requirement on forgiveness. Unless one has *antecedently* made up one's mind about the truth of both the objective impermissibility requirement and the objective blameworthiness requirement on forgiveness, the competing claim that their subjective rivals should replace both these putative requirements is *prima facie* reasonable. In addition, suppose we also accept principles *OIC* and *Equivalence*. Then the contender that forgiveness requires belief in impermissibility and not impermissibility *per se* is plausible as is the candidate that forgiveness requires belief in blameworthiness rather than blameworthiness *per se*.

Forgiveness and Determinism

The stage is now set to draw some conclusions about whether determinism imperils forgiveness. Take as unproblematic that determinism undermines freedom to do otherwise (See, e.g., Wiggins, 1973; van Inwagen, 1983; Ginet, 2003). Bridge principle *Blameworthiness / Impermissibility* gives us two alternatives regarding blameworthiness and impermissibility: either blameworthiness requires impermissibility (*BRI*) or belief in impermissibility (*BRBI*). Suppose *BRI* is true. Since *OIC* and *Equivalence* entail that nothing can be impermissible for one unless one could have done otherwise, and determinism precludes freedom to do otherwise, *prima facie*, nothing can be impermissible for anyone in a deterministic world.[1] Hence, if blameworthiness requires impermissibility, no one can be blameworthy for anything in a deterministic world.

Assume, next, that forgiveness requires both impermissibility and

1. I say "prima facie" because compatibilists may well wish to defend the view that that "can" of obligation is some weak compatibilist "can" and not some relatively strong incompatibilist "can." Compatibilists of this bent can accept the view that nothing can be impermissible or obligatory for you unless you could have done otherwise, and affirm that determinism does not preclude you from doing otherwise, given an apt compatibilist analysis of *one's being able to do otherwise*.

blameworthiness. To be forgiven, the offender must have done something morally wrong and for which she is to blame. Since deterministic worlds are shorn of impermissibility and blameworthiness, there is no fitting or apt forgiveness in such worlds. You may initially think that this is nothing momentous because it is the subjective member of each contrast pair that is true: necessarily, to be forgiven, the victim and offender believe that the offender has done wrong and believe that the offender takes herself to be blameworthy for this wrongdoing. Nothing precludes agents from having false beliefs in a deterministic world: although such a world is devoid of blameworthiness and wrongness, agents in this world may act on the basis of the false belief that they are doing wrong and falsely believe that they are blameworthy for their wrongdoing, and others may falsely believe that these agents have done wrong (or take themselves to have done wrong) and are to blame (or believe that they are to blame) for their wrongdoings. Hence, if the subjective members of either contrast pair are true, determinism does not preclude forgiveness.

However, even if the subjective members of the contrast pairs are to be preferred to their objective members, there is a bitter pill to swallow regarding forgiveness in deterministic worlds: there is a transparent sense in which judgments of blameworthiness, impermissibility, and forgiveness would be irrational in these worlds insofar as these judgments would be predicated on false beliefs, for example, the false belief that acts can be wrong for agents in these worlds. A practical corollary of this sobering conclusion is that if agents in deterministic worlds come to the realization, perhaps over time, that nothing they do is impermissible for them (or that everything they do is amoral for them), and that to be blameworthy for something in their deterministic world, they act on the false belief that they do wrong, no agent would be blameworthy for anything should agents give up these false beliefs. It would seem that knowing that nothing they do is impermissible for them, these agents (again, perhaps over time) would fail to act on the basis of the belief that they are doing wrong, and would fail to believe that agents in this world act on the basis of the belief that they are doing wrong. Hence, even if blameworthiness requires belief in impermissibility, it would be unrealistic to assume that these agents would ever be blameworthy for anything and that these agents would believe that anyone is blameworthy for anything. It follows that even if the subjective members of the contrast pairs are true, forgiveness would fade away in deterministic worlds.

Forgiveness and Artificial Agency

The wrongness contrast pair says that, necessarily, to be forgiven for something (i) it's wrong for the offender to do this, or (ii) the offender believes

that it's wrong for her to do this. The blameworthiness pair affirms that, necessarily, to be forgiven for something (i) the offender is blameworthy for this thing, or (ii) the offender believes that she's blameworthy for this thing. I have argued in favour of the second members – the belief or subjective members – of each pair. Regarding whether AI agents may be apt candidates of forgiveness, I remain agnostic about which member of each pair is true to guide the discussion.[1] The relevant issues to be tackled for what ensues are whether AI agents can do wrong or have relevant beliefs about wrongdoing, beliefs such as it is morally impermissible for me, R2D2, to make this decision, and whether AI agents can be blameworthy for at least some of their decisions or have the pertinent beliefs about being blameworthy. I assume (dogmatically) that AI agents are at least minimal agents – they can perform intentional actions, notably, mental actions of decision-making.

AI Agents and Wrongdoing

Sharif, an adorable Old English Sheepdog, performs many intentional actions (I take this to be uncontroversial) but does he ever do wrong? Compare: does a one-year-old human child ever do wrong? Does an adult chimpanzee ever do wrong? The challenging well-known dispute here concerns what may be dubbed "moral normative agency": what conditions must a minimal agent satisfy for at least some of her actions to instantiate the morally deontic properties of rightness, wrongness, or obligatoriness? The strategy I deploy to make some headway starts with the issue of "thick normative agency," which concerns relatively strong agency requirements of moral responsibility.

One such requirement is that the candidate be capable of intentional deliberative action. Such action, in turn, requires some psychological basis for evaluative reasoning. To be responsible, for example, one must, at least on occasion, be able to assess which line of conduct is best from one's own perspective of what is best. An agent's deliberations involve an appraisal of reasons for or against action. They issue in a practical judgment, which in turn gives rise to a decision or intention.

Drawing on previous work, I propose that an agent requires an *evaluative scheme* (or constituents of such a scheme) to entertain such reasons and to appraise them (see, e.g., Haji & Cuypers, 2008, ch. 3). An evaluative scheme has these constituents: (a) Normative standards the agent believes (though not necessarily consciously) ought to be invoked in assessing reasons for action, or beliefs about how she should go about making choices. Leave it open what

1. As previously noted, I leave it open that both members may be false.

these standards might be, or how much weight one gives to standards that, on occasion, deliver conflicting verdicts. To be a fitting candidate for *moral* responsibility, an agent's evaluative scheme's deliberative standards must include a set of *moral* principles or norms to ensure minimal moral competence: the agent must grasp the concepts of rightness, obligatoriness, or wrongness and be able to appraise, morally, various choices or actions in light of the moral norms that are elements of her evaluative scheme. There is no requirement that these moral appraisals be fully considered, free of error, or even conscious. Nor is there any requirement that these norms be evidentially based or justified in any strong sense of 'justification.' The agent may simply assimilate, without critical scrutiny, various norms of her religion or culture. (b) Included in the agent's evaluative scheme are desires, beliefs, or plans that express the agent's long-term ends or goals she deems worthwhile or valuable. For example, an agent may underscore her commitment to attempt to maximize overall happiness whenever she acts. (c) An evaluative scheme also has, as constituents, deliberative principles that the agent utilizes to arrive at practical judgments about what to do or how to act. For instance, the agent may endorse a principle of this sort: when self-interest conflicts with the demands of peer pressure, always side with self-interest. Or one may believe that the best way to maximize overall happiness is to rely on rules of thumb, such as, "keep your promises," "don't cheat," "don't steal," and so forth. (d) Lastly, an evaluative scheme incorporates motivation both to act on the normative standards specified in (a) and to pursue one's goals of the sort described in (b) at least partly on the basis of engaging the deliberative principles outlined in (c).

One may opt to refine (a), (b), (c), and (d). One might, for instance, wonder whether satisfaction of any of these conditions entails that the agent is self-conscious. Here, I rest (uneasily) with what has been documented about (a), (b), (c), and (d), fully recognizing that far more needs to be said to appreciate the complex notion of *being a normative agent*.

It *suffices* for an individual's being a normative moral agent at a certain time that the individual have at that time (i) an evaluative scheme with the requisite moral elements – the agent is minimally morally competent; (ii) deliberative skills and capacities; for example, the agent has the capacity to apply the normative standards that are elements of her evaluative scheme to assess reasons; and (iii) executive capacities – the agent is able to act on at least *some* of her intentions, decisions, or choices; she has the capacity to translate some of her intentions, decisions, or choices into action. An individual, like a toddler, who fails to have deliberative or executive capacities will be able to exert much less control, if any, over its actions than an

individual who does have such capacities. Interpret condition (ii) to entail that the agent can engage in *genuine* deliberation; her deliberative activities must meet the threshold of rationality below which such activities fail to count as bona fide *deliberation*. Persons suffering from certain severe psychological ailments may not be capable of genuine deliberation.

Normal, mentally healthy, adult human beings are normative moral agents. Sheep dogs, like Sharif, and human babies are not. Presumably, Sharif is not minimally morally competent. Not having moral standards, Sharif cannot apply any such standards to evaluating reasons for action even if we assume that he does undertake crude evaluations of his reasons. A plausible proposal is that an entity or agent can perform actions that are right, wrong, or obligatory only if it is a normative moral agent, and one is a normative moral agent only if one is minimally morally competent. On this moral competence model, to be a normative moral agent, one must have a basic grasp of the concepts of rightness, wrongness, and obligatoriness, and one must be able to use these concepts to appraise various things.

Although I am attracted to the moral competence model of deontic moral agency, I suspect that its requirements are too stringent: things that fail to be minimally morally competent may, it seems, perform actions that can be right, wrong, or obligatory. Think, for instance, of hypothetical agents who are devoid of moral concepts, do perform intentional actions, and undertake aesthetic or prudential evaluations of various things like policies, courses of action, paintings, architecture, and the like. Such beings would be normative agents of a particular variety: though normatively competent, their normative competence would not include moral competence. These beings would have an evaluative scheme with appropriate aesthetic and prudential elements but not moral elements, deliberative skills and capacities, and executive capacities. Intuitively, such beings may well perform actions that are morally wrong, although they wouldn't know that these were morally wrong, or they wouldn't perform them in light of the belief that they were doing moral wrong. With this sort of example in mind, one might offer, in contrast to the moral competence model, a normative competence model according to which an entity can perform actions that are right, wrong, or obligatory only if it is a *bona fide* normative agent of some sort; it must have an evaluative scheme that includes normative standards for evaluation, deliberative skills and capacities, and executive capacities. Such an entity would be able to undertake some sort of normative appraisal – one that needn't be moral – of things like policies, actions, and so forth.

It may be objected that whereas this relatively weak model gets the case of the hypothetical beings right, one of its drawbacks is that it rules that a tiger's act of, say, killing a human child, is wrong. But this is controversial. It is

debatable, for instance, whether tigers have anything like an evaluative scheme, even one that need not include moral standards of evaluation.

I shall not decide between the two models though, of the two, I'm drawn to the weaker second. Reverting to AI agents, and focusing, first, on whether they can do wrong, or act on the basis of the belief that they are doing wrong, one would need to ascertain whether AI agents possessing deliberative and executive capacities and equipped with evaluative schemes that include appropriate moral elements can be constructed. Speculating, I see no reason to deny that it is possible to fabricate such agents.

AI Agents and Blameworthiness

Can AI agents be built that are the sort of agent capable of commissions or omissions for which they can be morally blameworthy or, more generally, morally responsible? Could they or their relevant behavior satisfy the conditions of moral responsibility? A non-exhaustive list of these conditions includes: (1) A control or freedom condition. (2) An agency condition. (3) Many have proposed an ownership condition: roughly, the springs of action that issue in one's decisions must be "truly one's own" and not, for instance, the product of covert manipulation of which one is unaware and cannot thwart (see, e.g., Haji & Cuypers, 2008, chs. 2-3; Kane, 1996, ch. 5). (4) An epistemic requirement perhaps along the lines that the agent believes that she is doing wrong. (5) What may be dubbed a "moral requirement," alleged examples of which are the blameworthiness-requires-impermissibility condition or blameworthiness-requires-belief-in-impermissibility condition. I offer this list as an invitation for further discussion on whether anything prevents development of AI agents who meet these conditions. This task is monumental. To expose just a few of the challenges, there has been long-standing debate raging over each of these conditions. For instance, regarding control, does responsibility require reasons-responsiveness, being able to conform one's actions to what one values or to desires with which one identifies, or being able to do something other than what one does, given the same past and the laws of nature? As for the moral requirement, does blameworthiness – the negative mode of responsibility – require wrongdoing or merely belief in wrongdoing? Clarity on these conditions is simply the initial step. One would then have to ascertain whether it's possible for AI agents to satisfy these conditions. I very gladly leave these (and the requisite attendant) tasks to others.

▣ Conflict of Interest

The author declares no competing interests.

References

Allais, L. (2008). Wiping the slate clean: The heart of forgiveness. *Philosophy & Public Affairs, 36*(1), 33–68. https://doi.org/10.1111/j.1088-4963.2008.00123.x

Allais, L. (2014). Freedom and forgiveness. In D. Shoemaker & N. Tognazzini (Eds.), *Oxford studies in agency and responsibility* (Vol. 2, pp. 33–63). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198722120.003.0004

Arpaly, N. (2006). *Merit, meaning, and human bondage: An essay on free will*. Princeton University Press.

Belzer, M. (1998). Deontic logic. In *Routledge encyclopedia of philosophy*. Taylor & Francis. https://www.rep.routledge.com/articles/thematic/deontic-logic/v-1

Brink, D. (2021). *Fair opportunity, responsibility, and excuse*. Oxford University Press.

Campbell, J. K. (2011). *Free will*. Polity Press.

Clarke, R. (2015). Abilities to act. *Philosophy Compass, 10*(12), 893–904. https://doi.org/10.1111/phc3.12282

Copp, D. (1997). Defending the principle of alternate possibilities: Blameworthiness and moral responsibility. *Noûs, 31*(4), 441–456. https://doi.org/10.1111/0029-4624.00055

Copp, D. (2003). "Ought" implies "can": Blameworthiness and the principle of alternative possibilities. In D. Widerker & M. McKenna (Eds.), *Moral responsibility and alternative possibilities: Essays on the importance of alternative possibilities* (pp. 265–299). Ashgate.

Driver, J. (1992). The suberogatory. *Australasian Journal of Philosophy, 70*(3), 286–295. https://doi.org/10.1080/00048409212345181

Feldman, F. (1986). *Doing the best we can*. D. Reidel.

Fields, L. (1994). Moral beliefs and blameworthiness. *Philosophy, 69*(270), 397–415. https://doi.org/10.1017/S0031819100047239

Fischer, J. M. (2006). *My way: Essays on moral responsibility*. Oxford University Press.

Frankfurt, H. G. (1969). Alternate possibilities and moral responsibility. *Journal of Philosophy, 66*(23), 829–839. https://doi.org/10.2307/2023833

Frankfurt, H. G. (1999). On caring. In *Necessity, volition, and love* (pp. 155–180). Cambridge University Press.

Franklin, C. (2018). *A minimal libertarianism: Free will and the promise of reduction*. Oxford University Press.

Ginet, C. (2003). Libertarianism. In M. J. Loux & D. W. Zimmerman (Eds.), *The Oxford handbook of metaphysics* (pp. 252–271). Oxford University Press.

Haji, I. (1994). Changing obligations and immutable blameworthiness. *Theoria, 60*(1), 48–62. https://doi.org/10.1111/j.1755-2567.1994.tb00877.x

Haji, I. (2002). *Deontic morality and control*. Cambridge University Press.

Haji, I. (2006). Frankfurt examples, obligation, and responsibility. *Journal of Ethics, 10*(3), 255–281. https://doi.org/10.1007/s10892-005-5781-1

Haji, I. (2016). *Luck's mischief*. Oxford University Press.

Haji, I. (2019). *The obligation dilemma*. Oxford University Press.

Haji, I. (2023). *Obligation and responsibility*. Oxford University Press.

Haji, I., & Cuypers, S. (2008). *Moral responsibility, authenticity, and education*. Routledge.

Hanser, M. (2005). Permissibility and practical inference. *Ethics, 115*(3), 443–470. https://doi.org/10.1086/428457

Hebert, R. (2016). *Teleologism full stop: A general theory of ability, agency, know-how, obligation, and justification* [Doctoral dissertation, University of Calgary].

Hieronymi, P. (2001). Articulating an uncompromising forgiveness. *Philosophy and Phenomenological Research, 62*(3), 529–555. https://doi.org/10.1111/j.1933-1592.2001.tb00073.x

Hilpinen, R., & McNamara, P. (2013). Deontic logic: A historical survey and introduction. In D. M. Gabbay, J. F. Horty, X. Parent, R. van der Meyden, & L. van der Torre (Eds.), *Handbook of deontic logic and normative systems* (pp. 3–36). College Publications.

Holmgren, M. R. (1993). Forgiveness and the intrinsic value of persons. *American Philosophical Quarterly, 30*(4), 341–351.

Holmgren, M. R. (2012). *Forgiveness and retribution: Responding to wrongdoing*. Cambridge University Press.

Hughes, P. M. (1993). What is involved in forgiving? *Journal of Value Inquiry, 27*(3), 331–340. https://doi.org/10.1007/BF01087682

Kane, R. (1996). *The significance of free will*. Oxford University Press.

Littlejohn, C. (2012). Does "ought" still imply "can"? *Philosophia, 40*(4), 821–828. https://doi.org/10.1007/s11406-012-9366-5

McNamara, P. (1996). Making room for going beyond the call. *Mind, 105*(420), 415–450.

McNamara, P. (2019). Deontic logic. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2019 ed.). https://plato.stanford.edu/archives/sum2019/entries/logic-deontic/ (Original work published 2006)

Mellema, G. (1991). *Beyond the call of duty: Supererogation, obligation, and offence*. State University of New York Press.

Moore, G. E. (1912). *Ethics* (W. H. Shaw, Ed.). Oxford University Press.

Murphy, J. G., & Hampton, J. (1988). *Forgiveness and mercy*. Cambridge University Press.

Nelkin, D. K. (2013). Freedom and forgiveness. In I. Haji & J. Caouette (Eds.), *Free will and moral responsibility* (pp. 165–188). Cambridge Scholars Publishing.

Nelkin, D. K. (2011). *Making sense of freedom and responsibility*. Oxford University Press.

Parfit, D. (1984). *Reasons and persons*. Clarendon Press.

Pereboom, D. (2001). *Living without free will*. Cambridge University Press.

Pereboom, D. (2014). *Free will, agency, and meaning in life*. Oxford University Press.

Pereboom, D. (2021). *Wrongdoing and the moral emotions*. Oxford University Press.

Rönnedal, D. (2009). *An introduction to deontic logic*. CreateSpace Independent Publishing Platform.

Smith, H. (1991). Varieties of moral worth and moral credit. *Ethics, 101*(2), 279–303. https://doi.org/10.1086/293289

Thomson, J. J. (1991). Self-defense. *Philosophy & Public Affairs, 20*(4), 283–310.

van Inwagen, P. (1983). *An essay on free will*. Clarendon Press.

Vihvelin, K. (2004). Free will demystified: A dispositional account. *Philosophical Topics, 32*(1–2), 427–450. https://doi.org/10.5840/philtopics2004321/211

Vihvelin, K. (2013). *Causes, laws, and free will*. Oxford University Press.

Vranas, P. (2007). I ought, therefore I can. *Philosophical Studies, 136*(2), 167–216. https://doi/org/ 10.1007/s11098-007-9071-6

Warmke, B. (2016). The normative significance of forgiveness. *Australasian Journal of Philosophy, 94*(1), 1–17. https://doi.org/10.1080/00048402.2015.1126850

Warmke, B., & McKenna, M. (2013). Moral responsibility, forgiveness, and conversation. In I. Haji & J. Caouette (Eds.), *Free will and moral responsibility* (pp. 189–212). Cambridge Scholars Publishing.

Widerker, D. (1991). Frankfurt on "ought implies can" and alternative possibilities. *Analysis, 51*(4), 222–224. https://doi.org/10.1093/analys/51.4.222

Wiggins, D. (1973). Towards a reasonable libertarianism. In T. Honderich (Ed.), *Essays on freedom of action* (pp. 31–61). Routledge & Kegan Paul.

Zimmerman, M. J. (1996). *The concept of moral obligation*. Cambridge University Press.

Zimmerman, M. J. (1997). A plea for excuses. *American Philosophical Quarterly, 34*(2), 229–243.