

# script

When sorry is the hardest game  
to play



Arts & Humanities  
Research Council

# An intuition test

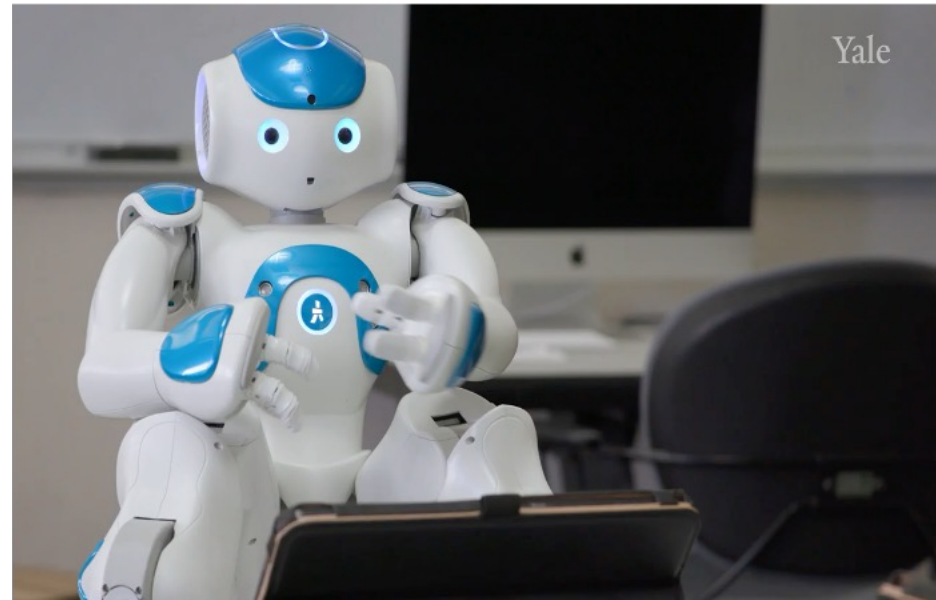


Arts & Humanities  
Research Council

script

# These may not be the apologies you are looking for

“Sorry, guys, I made the mistake this round,” it says. “I know it may be hard to believe, but robots make mistakes too.”



Arts & Humanities  
Research Council

script

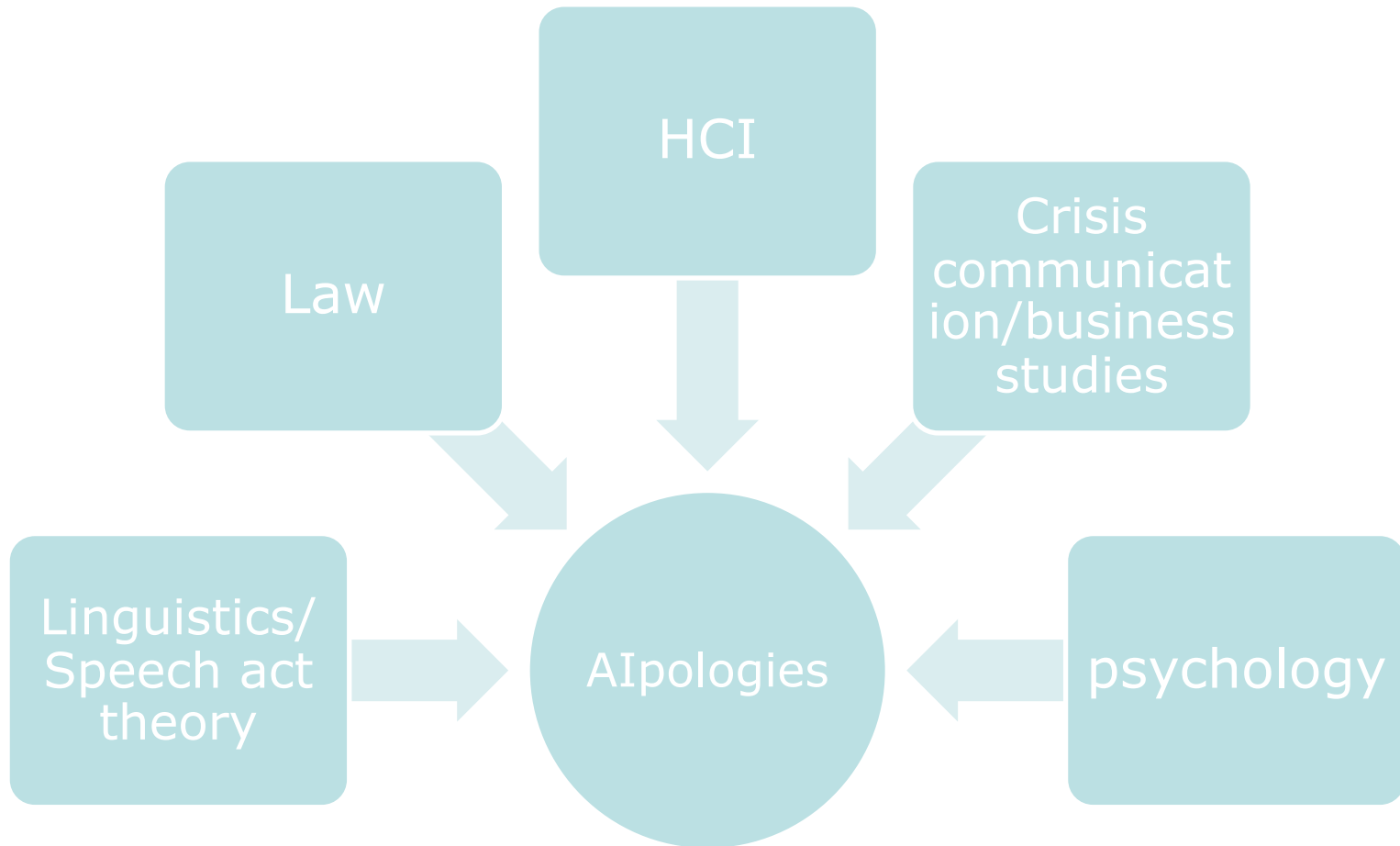
- This talk is about apologies by and from AIs
- - AIpologies so to speak
- I apologize for this terrible pun
- ...it won't be the only one though



Arts & Humanities  
Research Council

script

# The research field



# Human apologies and Trust

› [J Appl Psychol](#). 2004 Feb;89(1):104-18. doi: 10.1037/0021-9010.89.1.104.

## Removing the shadow of suspicion: the effects of apology versus denial for repairing competence-versus integrity-based trust violations

[Peter H Kim](#)<sup>1</sup>, [Donald L Ferrin](#), [Cecily D Cooper](#), [Kurt T Dirks](#)

Affiliations + expand

PMID: 14769123 DOI: [10.1037/0021-9010.89.1.104](#)

--



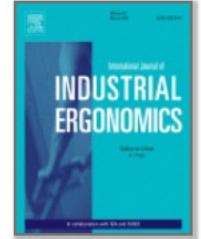
Arts & Humanities  
Research Council

script



# International Journal of Industrial Ergonomics

Volume 82, March 2021, 103078



## Robot apology as a post-accident trust-recovery control strategy in industrial human-robot interaction

Piotr Fraczak <sup>a</sup>  , Yee Mey Goh <sup>a</sup> , Peter Kinnell <sup>a</sup> , Laura Justham <sup>a</sup> , Andrea Soltoggio <sup>b</sup> 

Show more 

 Share  Cite



Arts & Humanities  
Research Council



# “I Don't Believe You”: Investigating the Effects of Robot Trust Violation and Repair

Publisher: IEEE

[Cite This](#)

[PDF](#)

Sarah Strohkorb Sebo ; Priyanka Krishnamurthi ; Brian Scassellati [All Authors](#)

41

Cites in  
Papers

1618

Full  
Text Views



# Evaluating the Impact of Emotional Apology on Human-Robot Trust

Publisher: IEEE

[Cite This](#)

[PDF](#)

Jin Xu ; Ayanna Howard [All Authors](#)

4

Cites in  
Papers

280

Full  
Text Views



Arts & Humanities  
Research Council

script



# What Makes An Apology More Effective? Exploring Anthropomorphism, Individual Differences, And Emotion In Human-Automation Trust Repair

Peggy Pei-Ying Lu, Makoto Konishi, Shin Sano, Sho Hiruta, Francis Ken Nakagawa

Recent advances in technology have allowed an automation system to recognize its errors and repair trust more actively than ever. While previous research has called for further studies of different human factors and design features, their effect on human-automation trust repair scenarios remains unknown, especially concerning emotions. This paper seeks to fill such gaps by investigating the impact of anthropomorphism, users' individual differences, and emotional responses on human-automation trust repair. Our experiment manipulated various types of trust violations and apology messages with different emotionally expressive anthropomorphic cues. While no significant effect from the different apology representations was found, our participants displayed polarizing attitudes toward the anthropomorphic cues. We also found that (1). some personality traits, such as openness and conscientiousness, negatively correlate with the effectiveness of the apology messages, and (2). a person's emotional response toward a trust violation positively correlates with the effectiveness of the apology messages.

Subjects: Human-Computer Interaction (cs.HC)



Arts & Humanities  
Research Council

script

# The Robot That Showed Remorse: Repairing Trust with a Genuine Apology

Publisher: IEEE

Cite This

PDF

Babiche L. Pompe ; Ella Velner ; Khiet P. Truong **All Authors**

**3**  
Cites in  
Papers

**210**  
Full  
Text Views



## Abstract

### Document Sections

- I. Introduction
- II. Related work
- III. Method

## Abstract:

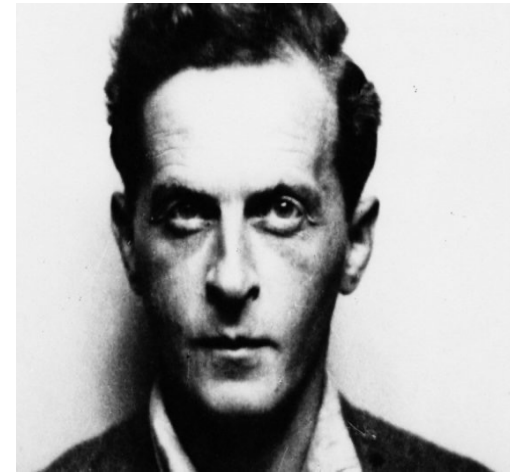
In the current state-of-the-art, robots are bound to make errors in a human-robot interaction (HRI). Trust is one of the important concepts in HRI that is often lowered by these errors. Fortunately, research has shown there are strategies that can help rebuild trust. An apology made by the robot is one of those strategies. However, apologies can take different forms. We designed a study in which Nao first built trust with the users, then violated that trust by making a speech recognition error, and then tried to restore it by either an apology with display of remorse, without remorse, or no apology at



Arts & Humanities  
Research Council



# UK-Austrian exchanges



- Ein philosophisches Problem hat die Form:
  - “Ich kenne mich nicht aus”



Arts & Humanities  
Research Council

script

- **420.** Aber kann ich mir nicht denken, die Menschen um mich her seien Automaten, haben kein Bewußtsein, wenn auch ihre Handlungsweise die gleiche ist wie immer? – Wenn ich mir's jetzt – allein in meinem Zimmer – vorstelle, sehe ich die Leute mit starrem Blick (etwa wie in Trance) ihren Verrichtungen nachgehen – die Idee ist vielleicht ein wenig unheimlich. Aber nun versuch einmal im gewöhnlichen Verkehr, z.B. auf der Straße, an dieser Idee festzuhalten! Sag dir etwa: »Die Kinder dort sind bloße Automaten; alle ihre Lebendigkeit ist bloß automatisch.« Und diese Worte werden dir entweder gänzlich nichtssagend werden; oder du wirst in dir etwa eine Art unheimliches Gefühl, oder dergleichen, erzeugen.



Arts & Humanities  
Research Council

script

- **420.** Aber kann ich mir nicht denken, die Menschen um mich her seien Automaten, haben kein Bewußtsein, wenn auch ihre Handlungsweise die gleiche ist wie immer? – Wenn ich mir's jetzt – allein in meinem Zimmer – vorstelle, sehe ich die Leute mit starrem Blick (etwa wie in Trance) ihren Verrichtungen nachgehen – die Idee ist vielleicht ein wenig **unheimlich**. Aber nun versuch einmal im gewöhnlichen Verkehr, z.B. auf der Straße, an dieser Idee festzuhalten! Sag dir etwa: »Die Kinder dort sind bloße Automaten; alle ihre Lebendigkeit ist bloß automatisch.« Und diese Worte werden dir entweder gänzlich nichtssagend werden; oder du wirst in dir etwa eine Art **unheimliches** Gefühl, oder dergleichen, erzeugen.



Arts & Humanities  
Research Council

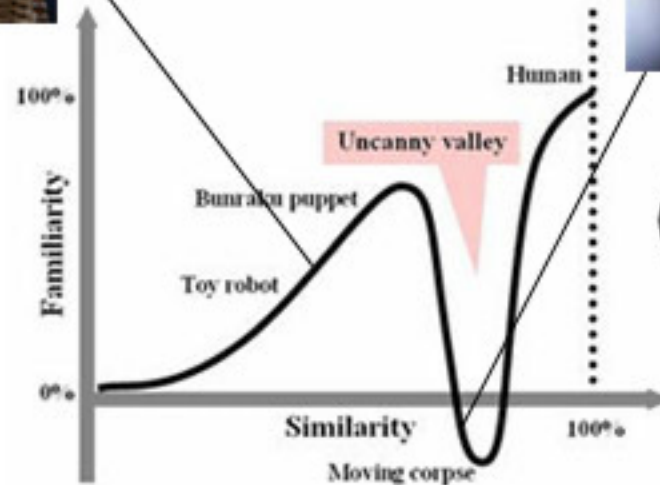
script

# The uncanny Otter Valley

(sorry, again...)



**CUTE**



**CREEPY**



Arts & Humanities  
Research Council

script

# Objection I: The curious case of the missing emotional state

- Philosophy of language: apology as expressives
- Criminal law: apology and criminal sanction



Arts & Humanities  
Research Council

script



# Sentencing Council

ANNUAL PUBLICATION, 2011

## Crown Court Sentencing Survey

24th May 2012

### Executive Summary

---

The Sentencing Council for England and Wales launched the Crown Court Sentencing Survey on 1 October 2010. It collects information directly from judges on the factors taken into account when they impose a sentence at the Crown Court.

This publication presents the findings of the survey for sentences passed by the Crown Court in 2011. The presentation of the results focuses on exploring the relationship between the factors taken into account in sentencing and the final sentence imposed, to help improve understanding of the sentencing process. The findings shown are across all offences, except where otherwise specified.



Arts & Humanities  
Research Council

script

- “For all offence types except robbery and sexual offences, genuine remorse on the part of the offender was the most common mitigating factor”



Arts & Humanities  
Research Council

script

# Apology and displays of pain

- “[the trial justice apparently detected no salt in the offender's tears; nor do we”
  - State v. Thornton, 800 A.2d 1016, 1045 (R.I. 2002)
- Bennet, Cristopher, The Apology ritual. A philosophical theory of punishment, (2008),



Arts & Humanities  
Research Council

script

# Show me those salty robot-tears

in Robotics and AI

[Front Robot AI](#). 2023; 10: 1121624.

PMCID: PMC10267379

Published online 2023 Jun 1. doi: [10.3389/frobt.2023.1121624](https://doi.org/10.3389/frobt.2023.1121624)

PMID: [37323644](https://pubmed.ncbi.nlm.nih.gov/37323644/)

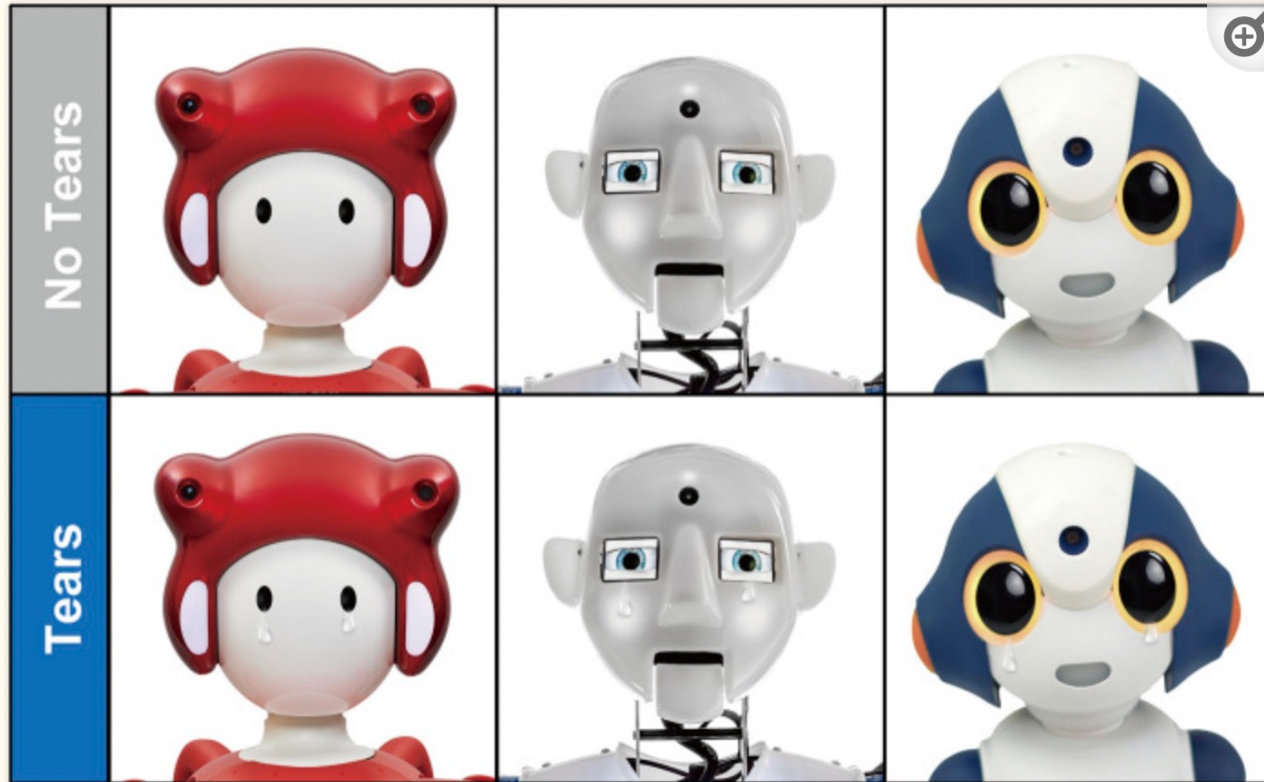
## Robots with tears can convey enhanced sadness and elicit support intentions

[Akiko Yasuhara](#)<sup>1,\*†</sup> and [Takuma Takehara](#)<sup>2,†</sup>

▶ [Author information](#) ▶ [Article notes](#) ▶ [Copyright and License information](#) [PMC Disclaimer](#)



# Show me those salty robot-tears



**FIGURE 1**

Visual stimuli used in the experiment. From the left, it is EMIEW, Robo Thespian, and Sota. The picture of RoBoHoN is not shown due to copyright issues.

# The ineffability of judgement



Arts & Humanities  
Research Council

script

 Restricted access | Research article | First published online June 2, 2015

## Affect and the Judicial Assessment of Offenders: Feeling and Judging Remorse


[Kate Rossmannith](#)  [View all authors and affiliations](#)

[Volume 21, Issue 2](#) | <https://doi.org/10.1177/1357034X14558073>

 Contents

 Get access

 Cite article

 Share options

 Information, rights and permissions



### Abstract

In most common law jurisdictions worldwide, an offender's remorse is a mitigating factor in sentencing. It matters whether or not a person who has committed a crime is truly sorry for what they have done. And yet how judges evaluate such expressions is unclear. Drawing on 18 interviews with judges in the New South Wales criminal justice system in Australia, this article examines the status of offenders' live, sworn evidence in the judiciary's assessment of offenders' remorse. These interviews with the judiciary reveal that remorse assessment often operates beyond semiotic, representational paradigms (such as 'demeanour assessment') and instead works, in experiential terms, as a feeling. When it comes to offenders getting into the witness box and speaking of their remorse, it seems that sometimes something gets felt by judges at the level of embodied affect that then enables them to declare: 'This person is remorseful.'



Arts & Humanities  
Research Council

Script

# The limits of robo-justice



“You cannot be just without being human”

Luc de Clapiers, marquis de Vauvenargues



This article from Netherlands Journal of Legal Philosophy is published by Boom juridisch and made available to anonieme bezoeker

## ARTICLES

# Sincere Apologies

## The Importance of the Offender's Guilt Feelings\*

*Margreet Luth-Morgan*

### 1 Introduction

The stronger position of the victim has led to new dynamics within the criminal justice process, such as the victim's statement of impact and alternative forms of dispute resolution such as victim-offender mediation. Within such a dynamic, the response from the offender becomes even more important. Ideal narratives of restorative justice refer to personal revelations brought on by the appeal on a

# Apology as punishment – an etymological play

- “Es tut mir leid”
- ich leide
- “I am sorry”
- “I am full of sorrow
- “am feel sore inside”.



Arts & Humanities  
Research Council

script

# Apologies, I promised some theology too and almost forgot

Legalistic remorse says, "I broke  
God's rules," while real repentance  
says, "I broke God's heart."

Timothy Keller

“ quotez fancy

 Download

 10 Wallpapers



 3  0



Arts & Humanities  
Research Council

script

# Danaher's ethical behaviourism

## Welcoming Robots into the Moral Circle: A Defence of Ethical Behaviourism

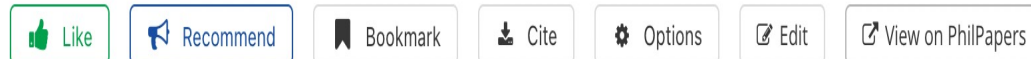
John Danaher



*Science and Engineering Ethics* 26 (4):2023-2049 (2020) [Copy](#) [BIBTeX](#)

### Abstract

Can robots have significant moral status? This is an emerging topic of debate among roboticists and ethicists. This paper makes three contributions to this debate. First, it presents a theory – ‘ethical behaviourism’ – which holds that robots can have significant moral status if they are roughly performatively equivalent to other entities that have significant moral status. This theory is then defended from seven objections. Second, taking this theoretical position onboard, it is argued that the performative threshold that robots need to cross in order to be afforded significant moral status may not be that high and that they may soon cross it (if they haven't done so already). Finally, the implications of this for our procreative duties to robots are considered, and it is argued that we may need to take seriously a duty of ‘procreative beneficence’ towards robots.



Arts & Humanities  
Research Council

script

# Danaher's ethical behaviourism

- (1) If a robot is roughly performatively equivalent to another entity whom, it is widely agreed, has significant moral status, then it is right and proper to afford the robot that same status.
- (2) Robots can be roughly performatively equivalent to other entities whom, it is widely agreed, have significant moral status.
- (3) Therefore, it can be right and proper to afford robots significant moral status



Arts & Humanities  
Research Council

script

# Wittgenstein on inner states of machines

- “Psychological language games ‘are as much a part of our natural history as walking, eating, drinking, playing’
  - Philosophical investigations para 25)
- Human language is ”full of soul”, whereas ‘the opposite of being full of soul is being mechanical’
  - *Remarks on the Philosophy of Psychology*
- ‘A machine thinks (perceives, wishes)’ seems somehow nonsensical. It is as though we had asked ‘Has the number 3 a colour?’
  - Black and Blue book at 47



Arts & Humanities  
Research Council

script

# *Philosophie der Psychologie. Band I,* §630.

- Statt des Unzerlegbaren, Spezifischen, undefinierbaren: die Tatsache, dass wir so und so handeln, z. B. gewisse Handlungen strafen, den Tatbestand so und so feststellen, Befehle geben, Berichte erstatten, Farben beschreiben, uns für die Gefühle der Anderen interessieren. Das Hinzunehmende, Gegebene – könnte man sagen – seien Tatsachen des Lebens / seien **Lebensformen**



Arts & Humanities  
Research Council

script

- 'Hurly-burly' and '*bustle*' of life as interpretative contexts
  - Remarks on the Philosophy of Psychology 625



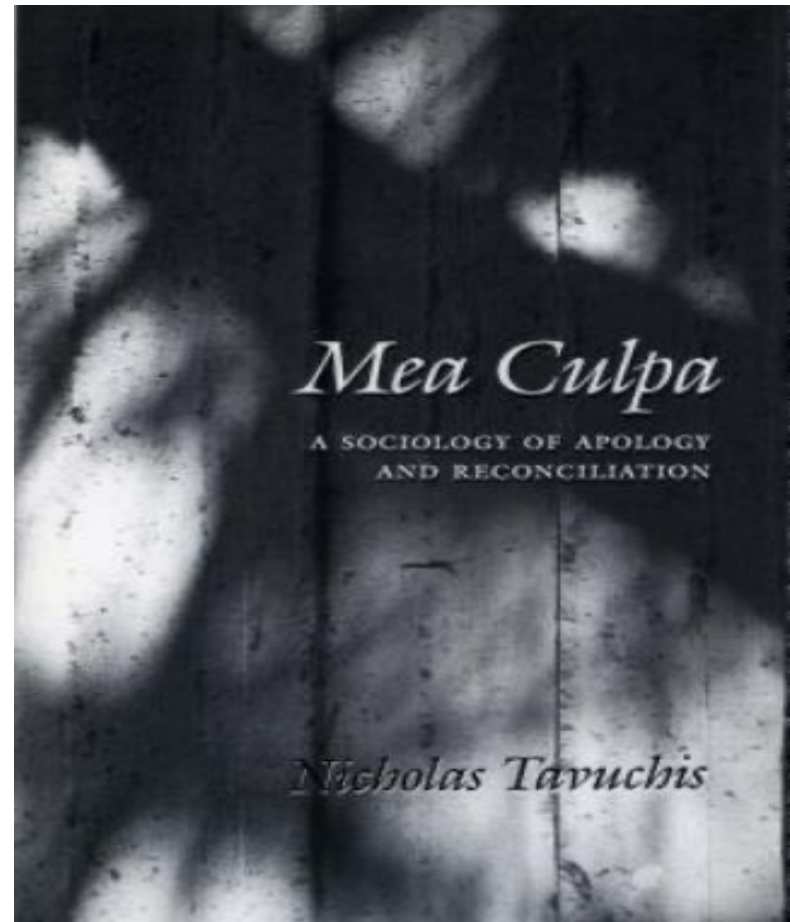
Arts & Humanities  
Research Council

script



# But do we have to think of apologies this way?

- an apology is a performative converts the remorse of the offender from "a private condition into public communion". The promise of change is so inextricable intertwined with the expression of remorse that it does not even need saying, it is always implied



Arts & Humanities  
Research Council

ot

# Apologies and business



## WE'RE SORRY

A chicken restaurant without any chicken. It's not ideal. Huge apologies to our customers, especially those who travelled out of their way to find we were closed. And endless thanks to our KFC team members and our franchise partners for working tirelessly to improve the situation. It's been a hell of a week, but we're making progress, and every day more and more fresh chicken is being delivered to our restaurants. Thank you for bearing with us.

Visit [kfc.co.uk/crossed-the-road](http://kfc.co.uk/crossed-the-road) for details about your local restaurant.

United CEO Oscar Munoz was scolded during a congressional hearing for neglecting to immediately apologize to the passenger who was forcibly removed from a flight in an incident that ignited public backlash.

"I was appalled at your comments at first," Congressman Lloyd Smucker, R-Pa., told Munoz during the Tuesday hearing.



Reuters/Kevin Lamarque

script



Arts & Humanities  
Research Council

# Apologies and nations

## Tony Blair apologies for Britain's role in the Slave Trade

15 March 2007

In a welcome move, Prime Minister Tony Blair said sorry for Britain's role in the Transatlantic Slave Trade.

*"I have said we are sorry and I say it again ... [It is important] to remember what happened in the past, to condemn it and say why it was entirely unacceptable,"* he said after meeting Ghana President John Agyekum Kufuor on 14 March.

Anti-Slavery International has been calling on the UK Government to make a formal apology for Britain's role in the Transatlantic Slave Trade and to take action to address its legacies, which continue to affect

### LATEST

- [Blog](#)
- [News](#)
- [Press Office](#)



Arts & Humanities  
Research Council

script

# Crisis management

Original Paper | [Published: 20 October 2018](#)

## The Value of Apology: How do Corporate Apologies Moderate the Stock Market Reaction to Non-Financial Corporate Crises?

[Marie Racine](#) , [Craig Wilson](#) & [Michael Wynes](#)

[Journal of Business Ethics](#) **163**, 485–505(2020) | [Cite this article](#)

**1031** Accesses | **4** Citations | **12** Altmetric | [Metrics](#)

Abstract



Arts & Humanities  
Research Council

script

# Aus gegebenem Anlass



ARTICLES

## Sorry, and not sorry, in Australia: how the apology to the stolen generations buried a history of genocide

Tony Barta

Pages 201-214 | Published online: 16 Jun 2008

Cite this article <https://doi.org/10.1080/14623520802065438>

Full Article

Figures & data

Citations

Metrics

Reprints & Permissions

View PDF



Arts & Humanities  
Research Council

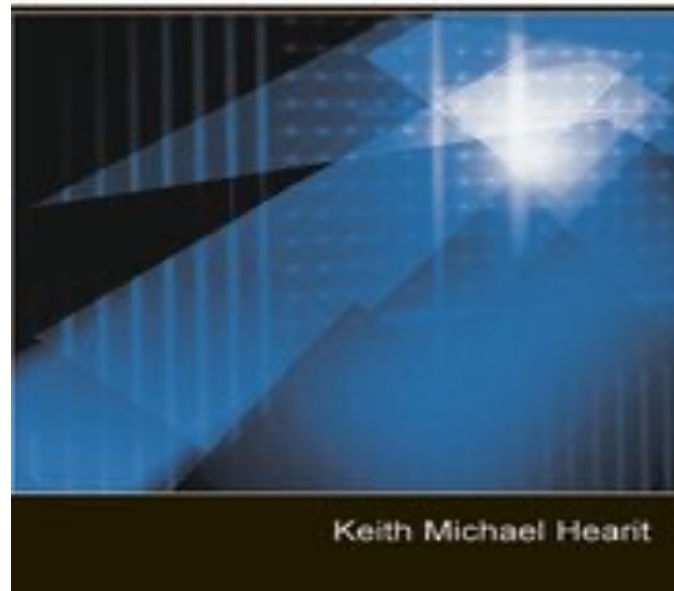
script

# Ethics of business apologies – the case for casuism

- Hearit “Apologetic ethic”

## Crisis Management by Apology

Corporate Responses to Allegations of Wrongdoing



script



Arts & Humanities  
Research Council

# From obligation to safe haven

[Home](#)

[About Us](#)

[Browse Legislation](#)

[New Legislation](#)

[Changes To Legislation](#)

[Search Legislation](#)

Title:  Year:  Number:  Type:

[Advanced Search](#)

## Apologies (Scotland) Act 2016

2016 asp 5 [Table of Contents](#)

[Table of Contents](#)

[Content](#)

[Explanatory Notes](#)

[More Resources](#)

[Plain View](#)

[Print Options](#)


### What Version

Latest available (Revised)

Original (As enacted)

### Opening Options

### More Resources

 [Original Print PDF](#)

[View more](#)

**Changes to legislation:** There are currently no known outstanding effects for the Apologies (Scotland) Act 2016.

#### Introductory Text

1. Effect of apology in legal proceedings
2. Legal proceedings covered
3. Definition of apology
4. No retrospective effect
5. Commencement
6. Short title

# Apologies as litigation protected statements if:

- States that the result is wrong
- "any statement made by or on behalf of a person which indicates that the person is sorry about, or regrets, an act, omission or outcome and includes any part of the statement which contains an undertaking to look at the circumstances giving rise to the act, omission or outcome with a view to preventing a recurrence".
- (gives a reason why)



Arts & Humanities  
Research Council

script



- an acknowledgment that there has been a bad outcome;
- an expression of regret, sorrow or sympathy for that bad outcome; and
- a recognition of direct or indirect responsibility for that bad outcome



Arts & Humanities  
Research Council

script

- Gives a reason why it won't happen again
- Gives an indication of a remedy



script



Arts & Humanities  
Research Council

# Do apologies have to be freely given?

› [Behav Sci Law](#). 2002;20(4):337-62. doi: 10.1002/bsl.495.

## Apology in the criminal justice setting: evidence for including apology as an additional component in the legal system

[Carrie J Petrucci](#) <sup>1</sup>

Affiliations + expand

PMID: 12210972 DOI: [10.1002/bsl.495](#)



Arts & Humanities  
Research Council

script

# Do apologies have to be freely given?

JOURNAL ARTICLE

## Against Court-Ordered Apologies

Nick Smith



*New Criminal Law Review: An International and Interdisciplinary Journal*

Vol. 16, No. 1 (Winter 2013), pp. 1-49 (49 pages)

Published by: University of California Press

[◀ Previous Item](#) | [Next Item ▶](#)



Arts & Humanities  
Research Council

script

# 'Drugs That Make You Feel Bad'? Remorse-Based Mitigation and Neurointerventions

Jonathan Pugh & Hannah Maslen



More download options

*Criminal Law and Philosophy* 11 (3):499-522 (2017) [Copy](#) [BIBTeX](#)

## Abstract

In many jurisdictions, an offender's remorse is considered to be a relevant factor to take into account in mitigation at sentencing. The growing philosophical interest in the use of neurointerventions in criminal justice raises an important question about such remorse-based mitigation: to what extent should technologically facilitated remorse be honoured such that it is permitted the same penal significance as standard instances of remorse? To motivate this question, we begin by sketching a tripartite account of remorse that distinguishes cognitive, affective and motivational elements of remorse. We then describe a number of neurointerventions that might plausibly be used to enhance abilities that are relevant to these different elements of remorse. Having described what we term the 'moral value' view of the justification of remorse-based mitigation, we then consider whether using neurointerventions to facilitate remorse would undermine its moral value, and thus make it inappropriate to honour such remorse in the criminal justice system. We respond to this question by claiming that the form of moral understanding that is incorporated into a genuinely remorseful response grounds remorse's moral value. In view of this claim, we conclude by arguing that neurointerventions need not undermine remorse's moral value on this approach, and that the remorse that such interventions might facilitate could also be authentic to the recipient of the neurointerventions that we discuss.



Arts & Humanities  
Research Council

script

# Wittgenstein

“The freedom of the will consists in the fact that future actions cannot be known now. We could only know them if causality were an inner necessity, like that of logical deduction.—The connection of knowledge and what is known is that of logical necessity”  
(TLP 1955: 5.136)



Arts & Humanities  
Research Council

script

# Maybe criminal law is not the best lens



Arts & Humanities  
Research Council

script

Never deny, never explain,  
never apologize



John Arbuthnot Fisher



Arts & Humanities  
Research Council





LORD FISHER TO THE RIGHT HON. WINSTON CHURCHILL.

MY DEAR WINSTON,

I AM here for a few days longer before rejoining my  
“Wise men” at Victory House—

“The World forgetting,  
By the World forgot!”

but some Headlines in the newspapers have utterly upset  
me! Terrible!!

“The German Fleet to assist the Land operations in  
the Baltic.”

“Landing the German Army South of Reval.”

We are five times stronger at Sea than our enemies  
and here is a small Fleet that we could gobble up in a few  
minutes playing the great vital Sea part of landing an  
Army in the enemies’ rear and probably capturing the  
Russian Capital by Sea!

This is “Holding the ring” with a vengeance!

Are we really incapable of a big Enterprise?

I hear that a new order of Knighthood is on the tapis  
—O.M.G. (Oh! My God!)—Shower it on the Ad-  
miralty!!

Yours,

FISHER.

9/9/17.



Arts & Humanities  
Research Council

script

- Recital 71 GDPR
- In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision.



Arts & Humanities  
Research Council

script

- Recital 71 GDPR
- In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, **to obtain an explanation of the decision reached after such assessment and to challenge the decision.**



Arts & Humanities  
Research Council

script

# Art 15 (h) GDPR

- the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.



Arts & Humanities  
Research Council

script

# EU HLEG AI

- “Explicability is crucial for building and maintaining users’ trust in AI systems. This means that processes need to be transparent, the capabilities and purpose of AI systems openly communicated, and decisions – to the extent possible – **explainable** to those directly and indirectly affected. Without such information, a decision cannot be duly contested.



Arts & Humanities  
Research Council

script

# EU AI ACT



Arts & Humanities  
Research Council

script

# TITLE 3 CH 2: System Requirements

- Art 13 - their **operation** is **sufficiently transparent** to enable users to interpret the system's output and use it appropriately ."



Arts & Humanities  
Research Council

script

# Recital 39

- Furthermore, the exercise of important procedural fundamental rights, such as the right to an effective remedy and to a fair trial as well as the right of defence and the presumption of innocence, could be hampered, in particular, where such AI systems are not sufficiently transparent, **explainable** and documented.



Arts & Humanities  
Research Council

script



*Harvard Journal of Law & Technology*  
Volume 31, Number 2 Spring 2018

**COUNTERFACTUAL EXPLANATIONS WITHOUT OPENING  
THE BLACK BOX: AUTOMATED DECISIONS AND THE GDPR**

*Sandra Wachter,\* Brent Mittelstadt,\*\* & Chris Russell\*\*\**

TABLE OF CONTENTS

I. INTRODUCTION .....	842
II. COUNTERFACTUALS .....	844
<i>A. Historic Context and the Problem of Knowledge</i> .....	846
<i>B. Explanations in A.I. and Machine Learning</i> .....	849
<i>C. Adversarial Perturbations and Counterfactual         Explanations</i> .....	851
<i>D. Causality and Fairness</i> .....	853
III. CONCLUSION .....	855



Arts & Humanities  
Research Council

script

- “In the existing literature, "explanation" typically refers to an attempt to convey the internal state or logic of an algorithm that leads to a decision. In contrast, counterfactuals describe a dependency on the external facts that led to that decision. This is a crucial distinction. In modern machine learning, the internal state of the algorithm can consist of millions of variables intricately connected in a large web of dependent behaviours”



Arts & Humanities  
Research Council

script

- “Giving a reason sometimes means “I actually went this way”, sometimes “I could have gone this way”, i.e. sometimes what we say acts as a justification, not as a report of what was done, e.g. I remember the answer to a question; when asked why I give this answer, I gave a process leading to it, though I didn’t go through this process.”
  - Wittgenstein, Lectures and Conversations on Aesthetics, Psychology, and Religious Belief (1966) at 22
- “The question “For what reasons do you believe this?” might mean: “From what reasons are you now deriving it (have you just derived it)?” But it might also mean: “With hindsight, what reasons can you give me for this supposition?”
  - Wittgenstein Philosophical Investigations (2009) 479



Arts & Humanities  
Research Council

script

# Explanation, hypotheticals and feedback

- IF you had done X (could do X in future) you get what you want
- IF you had £5.000 more, you'd have gotten the mortgage



# The AI is always right, right?

- *Assumes* as default that the decision was correct, AND that the addressee is in a position to bring about change
- Can at best give indirectly clues that a decision was badly wrong:

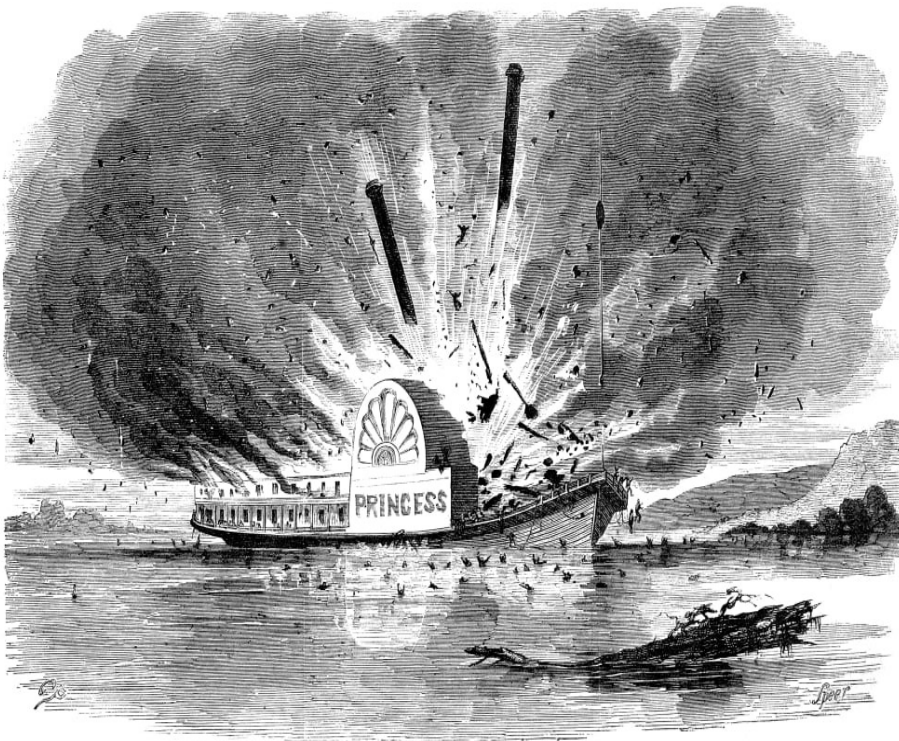


Arts & Humanities  
Research Council

script

# Science optimism as regulatory principle

- The rise of expert certification



# 2 mistakes

- AI: “If you earned more than £30000 annually, you would get the credit card”
- Customer: “But I do earn more than £30000 already, and said that much on section 8 of the form”
- More difficult is our example above, the use of an illegitimate criterion as opposed to a false fact:
  - AI: “If you had been male, you would have been given a credit card”
  - Customer: ”Hang on, that can’t be right...”



Arts & Humanities  
Research Council

script

- “The truly apocalyptic view of the world is that things do not repeat themselves. It isn't absurd, e.g., to believe that the age of science and technology is the beginning of the end for humanity; that the idea of great progress is a delusion, along with the idea that the truth will ultimately be known; that there is nothing good or desirable about scientific knowledge and that mankind, in seeking it, is falling into a trap. It is by no means obvious that this is not how things are



Arts & Humanities  
Research Council

script



# Apology as/instead of explanation

Inverts Wachter et al:

- “If YOU had done X, you would avoid Y”

to:

- “WE/I/AI should have done X to avoid Y”

AND

- We are sorry?
- It won't happen again?
- Have a drink on us?
- ....



Arts & Humanities  
Research Council

script

# Wittgenstein I

- TLP 6.421:

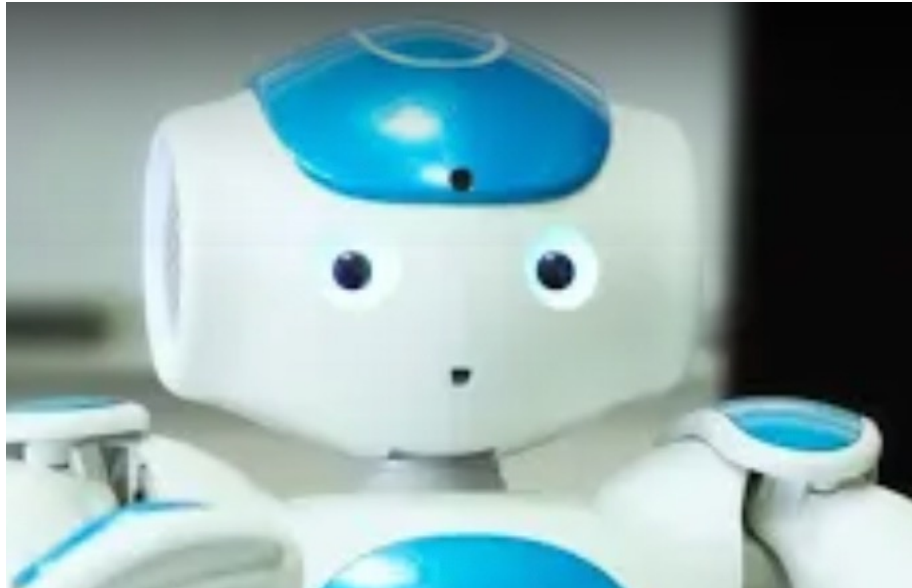
It is clear that ethics cannot be expressed.  
Ethics is transcendental.  
(Ethics and aesthetics are one.)



Arts & Humanities  
Research Council

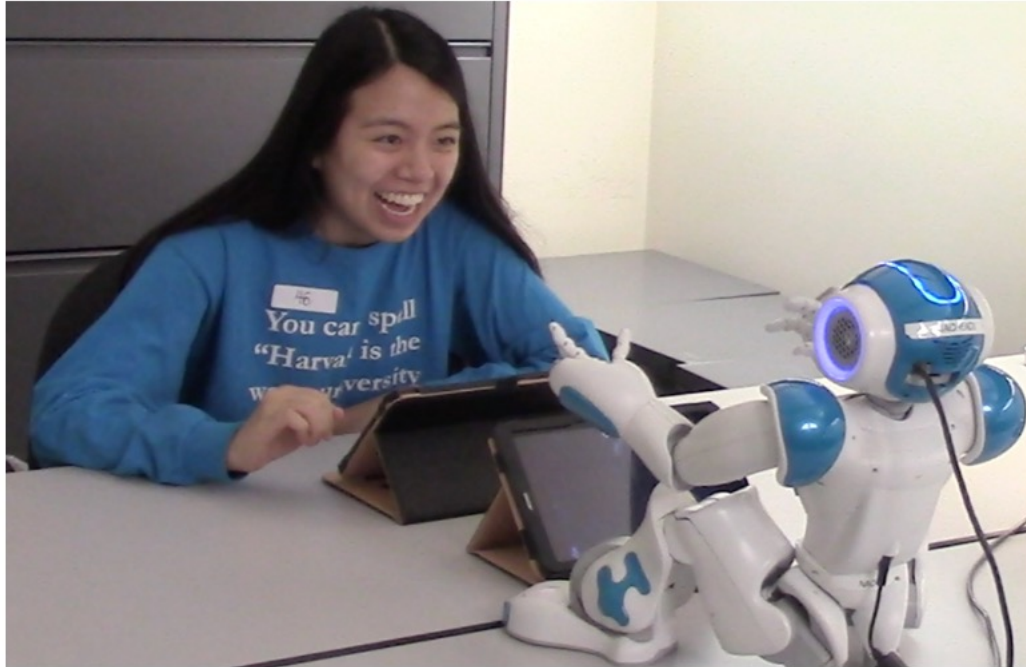
script

# Smile....



Arts & Humanities  
Research Council

*script*



script



Arts & Humanities  
Research Council

I]f one were trying to imagine a facial expression not susceptible of gradual and subtle alterations; but which had, say, just five positions; when it changed it would snap straight from one to another. Would this fixed smile really be a smile? And why not? – I might not be able to react as I do to a smile. Maybe it would not make me smile myself.

(RPP II, §614



Arts & Humanities  
Research Council

script



- What do you see?
- 'A face' (PI II, p. 204).
- best description I can give of what was shewn me for a moment' (PI II, p. 204
- drawing 'will be able straight away to reply to such questions as is it [...] "Smiling or sad?", etc.'
- (BBB, p. 163).



Arts & Humanities  
Research Council

script

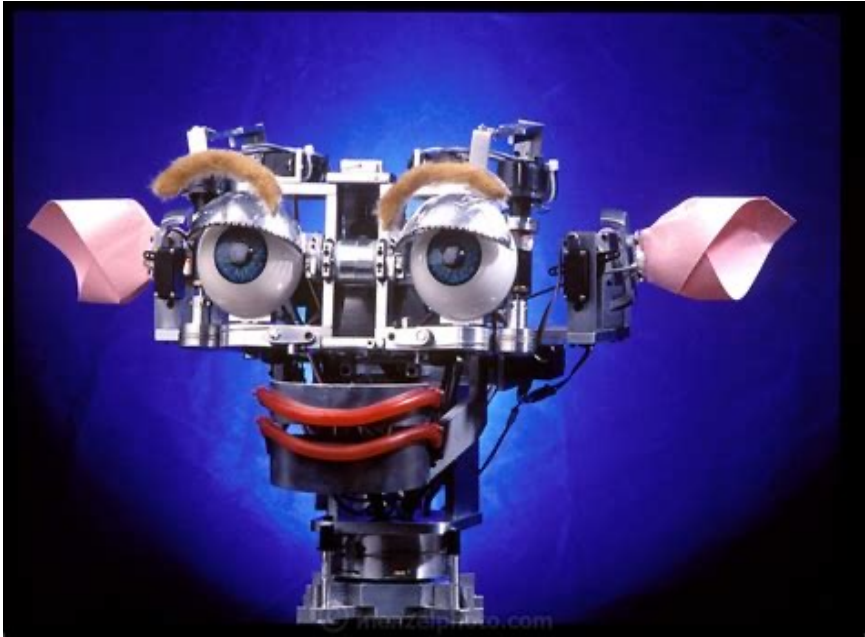
- 'In some respects I stand towards [the picture-face] as I do towards a human face. I can study its expression, can react to it as to the expression of the human face. A child can talk to picture-men or picture-animals, can treat them as it treats dolls' (PI II, p. 194).



Arts & Humanities  
Research Council

script

# mind-blind smiling-machine (Diane Proudfoot)





- What are the rules on refusing a (robot) apology?
- What are the ethical duties towards one who has apologized?



Arts & Humanities  
Research Council

script

# Relational theory of robot rights

## ROBOT RIGHTS

David J. Gunkel



script



Arts & Humanities  
Research Council

# Thinking the Unthinkable

Is/Ought Variations

Unthinkable



Cannot and  
Should Not



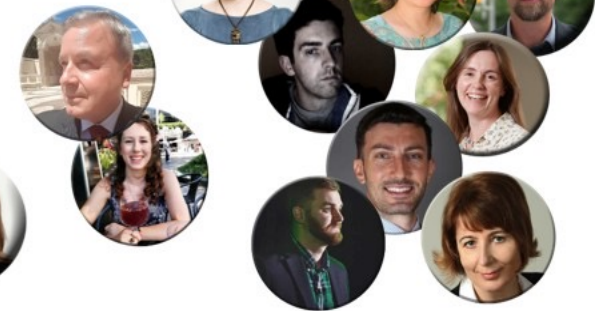
Can and Should

Thinking  
Otherwise



Can but  
Should Not

Cannot but  
Should



ROBOT RIGHTS NOW

# Wrapping up

- Aipologies *can* play a role in AI governance and also in generating trustworthy AI
- But only if
- We can distinguish proper and improper Aipologies on a conceptual level, and that distinction maps on what is technically feasible



Arts & Humanities  
Research Council

script

- A Wittgensteinian perspective may help us to distinguish proper and improper Apologies without resorting to mentalism
- A Wittgensteinian perspective may also help us to ground the respective duties and obligations between apologizer and apologisee without robot-right essentialism



Arts & Humanities  
Research Council

script

# The management apologizes for this talk

