Between Trust and Control

AIpology: when saying sorry is the hardest string to compute

Burkhard Schäfer

Apologies play an important role in trust recovery in post-conflict scenarios. As we increasingly interact with autonomous systems, HCI researchers too have discovered the power of apologies for situations where AIs or robots violated justified expectations of the humans they interact with. But are AIs the type of entity that can meaningfully apologise? Drawing on conceptions of apologies across a range of legal field, the chapter identifies requirements for robot-generated apologies that ensure not only their ethically sound deployment, but also, potentially, their recognition in law.

A. Introduction: The author wants to apologise for any inconvenience caused

This chapter explores how apologies generated by AIs – AIpologies – can generate, restore or sometimes undermine rational trust in autonomous devices, their ethical and legal implications, and what they can teach us more broadly about the intersection between trust, law and conflict.

At this point, I should apologise for the terrible AIpology pun – but also warn you that there are more to come. There are some other apologies I would like to make: I should apologise for some of the more challenging aspects of this chapter. It is located in the intersection of several disciplines: robotics, human-computer-interaction, psychology, business studies, linguistics, law, ethics and philosophy. As I can only claim expertise in a small sub-section of these, if any, my accounts may sometimes be wrong or misleading. If you don't understand any of the arguments, well then you'll have only your insufficient preparation to blame, and I recommend that you come back after doing some further reading. I tried initially to avoid this problem, and also save myself a lot of work, by simply having ChatGTP summarise the respective research fields and claim its insights as my own, unfortunately its output was spotted by the editors as machine generated, and I had to promise to write my own text.

If after reading the last paragraph, you now feel a mix of confusion, irritation, or even anger - then the rest of the chapter will hopefully be for you (and I apologise for leading you, for pedagogical purposes, briefly down this garden path). Apologies *can* be a powerful tool to restore trust after a norm violation occurred, and they also play an important role in several of the legal fields that ConTrust explored, from media regulation to criminal law to international law and post-conflict resolution between communities, societies and countries. As we will see, their potency has also been recognised increasingly in the field of robotics and human-computer interaction.

However, just in the same way in which we must distinguish trust from rational *trustworthiness*, we also have to distinguish the mere apology *rituals* from "rationally successful apologies". To fulfil their positive function, apologies have to be done the right way, and the above paragraph contained several violations of the felicity conditions for apologies as a specific kind of speech act. We will see how the difference between trust and justified trust, a distinction that was also central for ConTrust,[1] maps onto different types of AIpologies. While all of them potentially increase the feeling of trust in the recipients of the apology, only some of them can improve trustworthiness. A key question that we will have to explore is if AIs are at least in principle capable to generate not just sentences that contain the word "sorry" at the syntactically right place – a trivial task – but meet all the success conditions for valid apologising.

A second element that we can note in my attempted apologies is the close link between apologies and explanations. All but one of the "apologies" above contained also an "explanation" of sorts, though they differed in the explanans. One referred to my lack of skills and knowledge, the other to my lack of character, and we will delve a bit deeper into the linguistic and psychological research on apologies to explore the difference between these two below.

Explanations and explainable AI (XAI) have in recent years become a pivotal design requirement for law compliant autonomous software systems. One of the "apologies" offered above in particular shares some features with an influential approach to explainable AI, the counterfactual model proposed by Wachter, Mittelstadt and Russel (Henceforth WMR) in the context of the (contested) right to explanation for automated decision

[1] Rainer Forst, *The Justification of Trust in Conflict. Conceptual and Normative Groundwork*, (ConTrust Working Paper, No. 2, ConTrust 2022) 7 https://publikationen.ub.uni-frankfurt.de/frontdoor/index/index/docId/70591 <last accessed 1.5.2024>.

making in the GDPR.[2] This connection between apologies and explanations will allow us to ask if for legal purposes, where the law requires explanations or explainability, sometimes what it should be asking for are instead apologies, or conversely, whether an AI that can create valid apologies for its actions also complies with legal explainability requirements.

We will also see how WMR-type explanations *can* enhance trust, but work best in a collaborative environment where from the outset, both sides share a common goal. This too makes them the mirror image of apologies, whose explanatory force presumes, and is shaped by, conflict. This not only allows us to contrast explanations with excuses and apologies, it also creates a second conceptual link with ConTrust. ConTrust is premised on the insight that while traditionally, trust has been seen as juxtaposed to conflict, this overlooked the importance, but also fragility, of trust in conflict and post-conflict situations. Similarly, I will argue that some approaches to make AI trustworthy through explainability are premised on the same understanding of the relation between justification, transparency and trust, and not sufficiently responsive to the dynamical dimension where conflict and trust evolve in creative tension.

We can now introduce the three interrelated issues that this chapter hopes to address.

- Are autonomous machines the type of agent that can, in principle, make a trustworthiness- enhancing apology?
- If machines can apologise, what does this mean for AI regulation. Can they be treated also as a form of explanation where the law requires these? Should they get privileges for litigation purposes?

The next section will introduce and briefly discuss WMR's counterfactual model of explanation. It will conclude that while appropriate in many contexts, it can deliver inappropriate results in situations where apologies rather than explanations would be the expected response from a human interlocutor. We will then look at examples from HCI research that tries to give robots the ability to apologise to the humans they interact with. I will introduce briefly an experiment carried out by Institute for Network Science at Yale University, in which a mixed human-robot teams participated in a collaborative game. When the team lost, the robot would either stay

[2] Sandra Wachter, Brendt Mittelstadt, Chris Russell, 'Counterfactual explanations without opening the black box: Automated decisions and the GDPR' (2017) 31 Harv. Journal of Law & Technology 841.

silent, make a factual statement about the scores, or make a self-deprecating apology. The research showed the beneficial impact this last strategy had on group cohesion and trust repair, but for me also created a profound sense of unease. I will then try to account for this unease by discussing the way in which the law thinks about apologies, looking at examples across the case studies that were also at the centre of ConTrust: criminal law, media law, and political conflicts.

From this discussion, I will try to extrapolate those features that any legally relevant AI-generated apology should have. I will argue that to the extent that AIs are capable of meeting these requirements, their utterances should get appropriate legal recognition, too.

B. Better luck next time: the counterfactual approach to AI explanations

One important aspect of the current regulatory debate regarding AI is the demand for explainability. While at the beginning of the 21th century, George Orwell's *1984* encapsulated for many the fear of technology-enabled data *collection*, their increased *use* by powerful AI systems found another literary classic reference point. Kafka's *The Castle* anticipated the fear of the "Black Box Society",[3] where judgements are handed out by an impersonal machine whose inner workings are forever hidden from those affected, became a golden threat that tied together several regulatory initiatives.

The EU High Level expert group on AI for instance writes:

> "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."[4]

Transparency is seen here as precondition for trustworthiness: we trust those who are open with us.[5] It is also a precondition for agency: we can

[3] Frank Pasquale, *The black box society: The secret algorithms that control money and information* (Harvard University Press, 2015).

[4] EU High Level Expert Group on AI, '*Ethics Guidelines For Trustworthy AI*', (2019) https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai 13 <last accessed 1.5.2024>.

[5] Steven Norman, Bruce J Avolio, Fred Luthans, 'The Impact of Positivity and Transparency on Trust in Leaders and their Perceived Effectiveness' (2010) 3 The Leadership

often only decide how best respond to a decision when we understand the reasons on which it is based. Explainability then leads to transparency – once we know the reasons why a decision-maker decided against us, we can either agree with the reasoning and adjust our behaviour, or if we disagree with the reasoning, contest the decision.

One particularly influential proposal to turn legal requirements – at the time the (contested) explainability requirements of the GDPR – into actionable design decisions by software developers, is the above mentioned "counterfactual" approach by Wachter, Mittelstadt and Russel.

Consider an software agent that decides on credit card applications. After answering several set questions about the applicant, the system creates a risk model for them that combines data of past applicants and decisions about them with characteristics they share with the current applicant. Their risk score is then compared against a pre-defined value, and if the applicant is deemed too risky, the application is rejected.

A helpful explanation then could be of the form of a counterfactual: "Your application was rejected. But *if* your monthly income had been £50 higher, *then* application would have been granted."

WMR write about their approach:

> "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 modem machine learning, the internal state of the algorithm can consist of millions of variables intricately connected in a large web of dependent behaviours"[6]

This sees explanations, not as a mechanistic report of the inner workings of the decision maker, but as a chain of reasons from external facts to an utterance (the credit decision). We will follow this characterisation also in this paper.

One reason for this is that WMR's approach is much closer to the understanding of "explanation" that we find in law. When the law requires judges to "explain" their decisions, we are not normally looking for an autobiographical account ("I first got interested in justice when as a child…")

Quarterly 350. Applied to AI, see Warren J von Eschenbach, 'Transparency and the Black Box Problem: Why we do not Trust AI' (2021) 34 Philosophy & Technology 1607.

[6] Wachter and others (2) 845; in a similar vein, but with a more philosophically grounded analysis, John Zerilli. 'Explaining Machine Learning Decisions' (2022) 89 Philosophy of Science 1.

or as an account of the neurological basis of the movement of their mouth that uttered the decision ("the information presented by the prosecution triggered a c-fibre in my brain that led to a movement of my..."), even though these can be valid explanations for some purpose in some contexts.

Second, this understanding of the nature of an explanation also means we can speak of an AI explaining itself, without having to commit ourselves to talk about inner mental states, something that will become important when we distinguish different ways to conceptualise apologies.

WMR has been highly influential in the discussion on machine generated explanation. Counterfactual explanations have a number of desirable formal characteristics that make it possible for the AI to generate not only a number of them for any given decision, but also to rank them, recommending for instance the course of actions that is the least complicated for the applicant. A good explanation tells them to increase their savings by £50 every month for a year, rather than to go back to university, get a degree, and on that basis get a much higher paid job. While both can be strategies that achieve the desired result, the more outlandish they are, the more likely the applicant will not perceive them as guidance for action, but will feel mocked.

While technological feasibility will have undoubtedly contributed to the popularity of this approach, we can also speculate that it resonates in many ways with academics: a WMR explanation shares many aspects with good student feedback – not (merely) justifying a mark, but pointing to the ways in which it can be improved the next time round.

While this is an advantage in many contexts, in others it is either not applicable, or even harmful and counterproductive. The benefits are obviously greatest when there is recurrent engagement, less so for one-off interactions, just as students benefit most from feedback in their first essays, least in their final dissertation.

A somewhat different question is if the recommendation must be "actionable", that is if the addressee of the explanation must have it in their power to bring about the suggested change. Telling a credit card applicant "if you had been born to very rich parents, your application would have been successful" is not very helpful, true as it may be. Sometimes though, non-actionable explanations can be both appropriate and helpful – they tell the decision subject that there is nothing they can do, which can prevent self-blame or futile resource allocation, for instance if an application for a high-risk profession such as pilot is rejected due to a congenital illness that makes them prone of suffering brain embolisms at high altitudes.

The main benefit of a counterfactual explanation approach is that it assists the persons affected by an AI decision to react constructively, and through their actions change the outcome in the future. In cases where the AI decision was correct, this is extremely helpful. It is even more helpful when the interest of the AI provider and the subject of the decision ultimately converge. In the credit card example, while a rejection will feel painful, ultimately it is also in the interest of the applicant not to be burdened with a loan that they have no chance of repaying. Failing students who lack the required competency levels not only protects, in the case of law students and medics at least, the general public, but also them, from the stress that comes from being an "imposter" in high-stake environments to possible litigation against them for malpractice.

This discussion allows us to connect our discussion more directly with ConTrust. Counterfactual explanations work best in cooperative environments where there is a high level of background trusts between the parties and also the possibility of ongoing, mutually beneficial interactions between them. The responsible lender, the good teacher, or even the judge who does not want to see the accused before them again will give explanations of this type and their credibility also depends to a degree that the subject of the decision ultimately trusts in the benevolence of the decision maker. Just as ConTrust asked the question of the role of trust and trustworthiness in conflict situations, we can now also explore the limits of counterfactual explanations in conflict situations.

In some conflict situations, back-engineering the explanation could lead to undesirable actions. If for instance a money-laundering detection system refuses a transaction, it should not generate as an explanation: "The law requires that transactions above £10000 must not be anonymous. If the transaction had been split into two transactions of £5000 send a few hours apart, these transactions would have been approved". This problem has also been recognised in the EU AI Act, which exempts in Art 61 police users of AI from disclosing certain sensitive operational information even if it is needed by the developers of the system to assess if it is working correctly.

But even in less obviously adversarial scenarios, one objective that legislators pursue through a legally mandated use of explainable AI is to also to create contestability of results. Contestability is a corner stone of the rule of law and is irreconcilable with a black box society where "the computer says no" ends the discussion.[7] XAI therefore also need to cover situations where the AI comes to the wrong result, or where the situation between the parties is shaped by conflict rather than convergence of interest. Here counterfactual explanations can be highly inappropriate. We saw this already in the first paragraph, when I admitted to difficulties in explaining my ideas clearly, but then counterfactually explained that if *you* were to read up on the material, you would get more out of this chapter. But obviously, shifting the "duty to rectify" to you when the fault was all mine rendered this ineffectual as an apology, and indeed offensive.

Counterfactual explanations can assist contestability, but only indirectly. There are two ways how this can happen:

If the generated explanation refers to a false statement about the world as explanans, contestation is the most straightforward:

> 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"

This, strictly speaking is not an explanation at all, merely an attempt at one. More difficult is a situation where an illegitimate criterion as opposed to a false fact is given as part of the explanation:

> AI: "If you had been male, you would have been given a credit card"
> Customer: "Hang on, that can't be right..."

This may well be a "correct" explanation, in the sense that it faithfully describes how the AI reached its decision, and we may even grant for the sake of the argument that there is a relevant causal connection between gender and ability to repay credit. The explanation fails for legal reasons (and that means, fails in some, but not necessarily all, jurisdictions), because it uses an illegitimate explanans. In either case though, the applicant has to deduce that something went wrong – the AI is good at judging the applicant and telling them how to do it right, less good and helpful at judging itself. This can create significant burdens on the individual, especially in the second scenario that requires from the applicant knowledge and understanding of

[7] Margot E Kaminski, Jennifer M. Urban, 'The Right to Contest AI' (2021) 121 Columbia Law Review 1957; Marco Almada, 'Human Intervention in Automated Decision-making: Toward the Construction of Contestable Systems' In Floris Bex (ed), *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law,* (ACM 2019).

discrimination law, and the resources (in time, money etc) to take appropriate action.

This also highlights another way in which the counterfactual approach to explanation maps onto the "conflict-antagonistic" understanding of trust, trustworthiness and transparency that ConTrust challenged. It reflects an underlying trust in technology, which in turn shapes the understanding of the role of law regarding its governance. This leads to the paradox that even though the aim is to reign in technologies that are perceived as dangerous or even out of control, the method of control is rooted in the same optimism regarding our ability to predict, and with that control, our environment that gave rise to these technologies in the first place.

In the case of machine generated explanations, the paradox becomes particularly visible: If I require an explanation to trust the AI, why should I trust the AI to have generated a correct explanation? Maybe we need explainable AI, to be able to trust *that* module too. This is not facetious. Some of the more technically oriented criticism of WMR and other post-hoc explanations showed their vulnerability to both intentional and unintentional manipulation.[8] This means a user of an XAI system needs to understand its limitations and risks to make informed decisions how much they can trust the explanation that was given. Here too we find the tension between transparency and conflict – adversarial settings lend themselves particularly to the manipulation of the explanation module.[9] To assure the subject of a decision could therefore also require explaining the way the explanation was generated, and equally, the requirements of Art 14 of the EU AI Act that deal with the knowledge of training of the human in the loop may require an understanding of XAI in addition of understanding the logic that lead to the primary decision.

The counterfactual explanation model works best when the AI is right, and it is then up to the individual to adjust their actions to achieve a

[8] Dylan Slack and others, 'Counterfactual Explanations can be Manipulated' in Marc'Aurelio Ranzato and others (eds), *Advances in Neural Information Processing Systems 34* (NeurIPS 2001); Ahmad-Reza Ehyaei, and others, 'Robustness Implies Fairness in Causal Algorithmic Recourse' In Sara Fox and others (eds), *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency,* (ACM 2023).

[9] Dylan Slack and others, 'Fooling Lime and Shap: Adversarial Attacks on Post Hoc Explanation Methods', in Anette Markham and others, *Proceedings of the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society,* (ACM 2020); Sebastian Bordt and others, 'Post-hoc Explanations Fail to Achieve their Purpose in Adversarial Contexts' In Charles Isbell and others (eds), *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, (ACM 2022).

desired goal. Contestability in case the AI got it wrong is at best a side result, assumes knowledge by the affected individual to interpret the answer correctly, and requires that they use their resources to complain and contest the decision.

We can now ask how an approach to explainability would look like that takes the scenario where the AI did *not* decide correctly as a starting point. Rather than focussing on *creating* trust, how can we restore trust once it was broken?

Let us reconsider the two mini-dialogues from above:

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"
AI: "If you had been male, you would have been given a credit card"
Customer: "Hang on, that can't be right..."

In this situation, the decision maker committed a mistake, trust in them is now broken and needs to be repaired. How would a human act in this situation? One obvious and very natural trust repair strategy would be to apologise: "I am so sorry, I misremembered what you told me", or "You are quite right, I apologise, this can't be right, of course you get a credit card, and the first two months are on us". Structurally, apologies are the mirror image of WMR's counterfactual explanation. At a bare minimum, WMR's:

"If *YOU* had done/are going to do X, you would/will avoid Y"

Now becomes:

"*WE* should have done X to avoid Y, and [....]"

[...] stands for now as a placeholder that completes the apology. We will discuss some candidates for this below.

In the next section we will look at a real-world example of a robot apologising, to tease out some of the intuitions that will influence the answer to these questions.

C. *"Everything is my fault, I'll take the blame"*

That apologies can be highly effective in restoring trust when issued by humans is a well-supported fact, with a wealth of empirical studies from psychology showing the beneficial effects for trust repair.[10] Business psychology and management studies in particular have embraced for a long time the benefits of apologies for efficient leadership internally, and repair of trust with customers and the wider public externally.[11]

Their effectiveness has more recently been recognised also by HCI researchers and roboticists, and it seems indeed that apologies issues by a robot or chatbot can have the same positive effect on trust repair as those done by human interlocutors.[12] Industrial robots apologising for sudden unexpected movements improved post-incident trust in the human co-workers.[13] Two robots apologising for the same mistake increased customer trust in a service robot environment.[14] Even in high stake environments such as simulated emergency evacuation, a timely apology by the guide-robot helped repair trust in its abilities.[15]

However, *why* apologies are trust-enhancing, and furthermore, if they also enhance trustworthiness, is much more debatable. Some apologies are obviously superfluous, for instance apologising for bad weather, yet they still increase trust in the apologiser, human or machine.[16] Conversely,

[10] See e.g. Fengling Ma and others, 'Apologies Repair Trust via Perceived Trustworthiness and Negative Emotions, (2019) 10 Frontiers in Psychology 758; Aaron Lazare, *On Apology*. (Oxford University Press 2005); Chris Reinders Folmer, and others, 'Repairing Trust Between Individuals and Groups: The Effectiveness of Apologies in Interpersonal and Intergroup Contexts', (2021) 34 International Review of Social Psychology 14.

[11] See e.g. Eric Schniter, Roman M. Sheremeta, Daniel Sznycer, 'Building and Rebuilding Trust with Promises and Apologies', (2013) 94 Journal of Economic Behavior & Organization 242; Marie Racine, Craig Wilson, and Michael Wynes, 'The Value of Apology: How do Corporate Apologies Moderate the Stock Market Reaction to Non-financial Corporate Crises?', (2018) 163 Journal of Business Ethics 485; Wei Shao and others, 'Toward a theory of corporate apology: mechanisms, contingencies, and strategies', (2022) 56 European Journal of Marketing 3418.

[12] See e.g. Gyounghwa Na, Junho Choi, Hyunmin Kang, 'It's not my Fault, But I'm to Blame', (2023) International Journal of Human–Computer Interaction [2023] 1.

[13] Piotr Fratczak, and others, 'Robot Apology as a Post-accident Trust-recovery Control Strategy in Industrial Human-robot Interaction', (2021) 2 International Journal of Industrial Ergonomics 103078.

[14] Yuka Okada, and others, 'Two is Better than One: Apologies from two Robots are Preferred', 18 (2023) *PLOS one* https://doi.org/10.1371/journal.pone.0281604.

[15] Xinyi Zhang and others, 'Sorry, it was my Fault": Repairing Trust in Human-Robot Interactions' (2023) 175 International Journal of Human-Computer Studies 1.

[16] Alison Wood Brooks, Hengchen Dai, Maurice E. Schweitzer, ''I'm Sorry About the rain! Superfluous Apologies Demonstrate Empathic Concern and Increase Trust', (2014) 5 Psychological and Personality Science, 467.

not all apologies are equally efficient. For both humans and robots, apologies that reference competence deficits are more effective than those that reference character deficits.[17] If you re-read the introductory section, ask yourself if my apology for (almost) plagiarising with ChatGPT was as trust-restoring as my apology for straying into fields for which I lack training. And even more puzzling, apologies increase trust even in situations where people distrust the sincerity of the apology.[18] This points us to an important distinction that will concern us for the rest of this chapter:

a) Which, if any, type of apology by humans is rationally restoring violated trust?
b) Are robots capable in principle to produce the type of apology which, had it been given by a human, would rationally restore violated trust?

To unpack these questions, we will now look in more detail at one particularly interesting study into robot apologies. In 2018, researchers at Yale conducted an experiment in which a vaguely humanoid, child-sized robot played together with several humans in a group activity.[19] In order to win, all group members had to work together. What was tested was the effect that apologies by the robot after a lost game would have on the group. To engineer this, the robot would randomly fail at its task. In some groups, the robot would say nothing when its action caused the team to fail, in others it would make a mere factual statement (announcing the score), and in the third group it would apologise to the other players and display vulnerability:

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

Or

> "Sorry, I sometimes run out of memory and can't process things fast enough".

[17] Zhang, X., Lee, S.K., Maeng, H. and Hahn, S., 2023. Effects of Failure Types on Trust Repairs in Human–Robot Interactions. International Journal of Social Robotics, 15(9), pp.1619-1635.

[18] Alice MacLachlan, 'Trust me, I'm Sorry": The Paradox of Public Apology' 98 (2015) The Monist, 441.

[19] Sarah Strohkorb Sebo and others, 'The Ripple Effects of Vulnerability: The Effects of a Robot's Vulnerable Behavior on Trust in Human-robot Teams' In Takayuki Kanda, Selma Ŝabanović (eds), *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, (ACM 2018).

These interventions positively influenced the trust group members placed in the robot. They interacted more with the machine, showed greater willingness to listen to it, and also used more non-verbal cues of trust such as gaze that are typical for human-to-human communication.[20] According to the researchers, the robot accepted responsibility, and through self-disclosure made itself vulnerable. As vulnerability and trust are closely aligned concepts, if this is indeed what the robot is doing, the effect on the trust relations should not be surprising. We might wonder however what it means for a robot to "make itself vulnerable", and if this is a correct description of its actions. The Yale experiment is not the only one that frames robot apologies in the terminology of "vulnerability", and its findings align with other studies that tested human reactions to robot apologies.[21] But they also point to an obvious problem with this approach. John Wayne in *She Wore a Yellow Ribbon* famously said, "Never apologise Mister, it's a sign of weakness", and undoubtedly, for many humans apologising, or admitting mistakes, does come with a strong feeling of dread. But does the same apply in a meaningful way to a machine, or are robots that apologise merely deceiving their human collaborators, making these, rather than themselves, vulnerable? It has indeed been argued that there is something profoundly unethical and deceptive about robot apologies.[22] But is this a problem with robot apologies, specifically, or do they merely inherit the problematic and highly ambivalent features that all apologies as trust-recovery strategy exhibit? To answer this question, I suggest to analyse the way in which the law thinks about and uses apologies as a comparator.

[20] Margaret Traeger and others, 'Vulnerable Robots Positively Shape Human Conversational Dynamics in a human–robot team." (2020) 217 PNAS 6370.

[21] See e.g Nikolas Martelaro and others. 'Tell me More. Designing hri to Encourage more Trust, Disclosure, and Companionship." In Christoph Bartneck and others (eds), *11th ACM/IEEE International Conference on Human-Robot Interaction*, (IEEE 2016); for mutual vulnerability Zachary Daus, 'Designing Mutually Vulnerable Human-Robot Interaction: Challenges and Possibilities' (2021) 2 Giornale di Filosofia 127.

[22] Makoto Kureha, 'On the Moral Permissibility of Robot Apologies' (2023) 38 AI & Society, 1.

D. Those salty robot tears

The most obvious objection against the Yale robot is that its statements are, in essence, lies. Apologies in this view are also, or even mainly, reports of an inner state. In Searle's terminology, they are expressives.[23] To be sincere, they require a feeling of remorse or regret. "Whatever else is said or conveyed, an apology must express sorrow".[24] Robots lack internal emotional states, so they cannot possibly truthfully apologise. The robot *says* sorry, but it *is* not sorry.

Other studies in machine apologies went in this respect much further than the Yale experiment. A particularly problematic example is the experiment by Pompe et al, that showed that explicit expressions of remorse are particularly effective in restoring trust.[25] To achieve what they call "genuine apology", they combine the verbal expression of remorse with appropriate body language, using the same type of vaguely anthropomorphic machine. *If* a feeling of remorse however is a defining element of true apologies, then this is merely an even more devious form of manipulation.

In legal contexts, displays of remorse are particularly important during sentencing in criminal trials.[26] Displays of sincere remorse is seen as a redeeming quality that merits consideration, while lack of remorse is seen as an indication of dangerousness.[27] Remorse and apology become here an indicator if not of "good character", then at least of "character capable of redemption". This requires more than an abstract, "learned" recognition of one's wrongdoing, rather, what sways judges and juries is evidence of

[23] John R Searle, 'A Classification of Illocutionary Acts.', 5 (1976) *Language in society* 1, 4 and 12-17.

[24] Nicholas Tavuchis, *Mea Culpa: A Sociology of Apology and Reconciliation*, (Stanford University Press 1993) 36.

[25] Babiche L. Pompe, Ella Velner, Khiet P. Truong 'The Robot that Showed Remorse: Repairing Trust with a Genuine Apology.' In Silvia Rossi & Antonio Sgorbissa (eds) 31st *IEEE International Conference on Robot and Human Interactive Communication RO-MAN)*, (IEEE 2022).

[26] So in particular Cristopher Bennet, *The Apology Ritual. A Philosophical Theory of Punishment*, (Cambridge University Press 2008).

[27] A list of examples, with an ultimately sceptical assessment, is In Jeffrie G. Murphy, 'Remorse, Apology, and Mercy' (2006) 4 Ohio St. J. Crim. L. 423; see also Stephanos Bibas, Richard A. Bierschbach, 'Integrating Remorse and Apology into Criminal Procedure.' (2004) 114 Yale lJ 85.

almost physical pain, expressed e.g. by tears.[28] As the Court put it in State v Thornton: "[the trial justice apparently detected no salt in the offender's tears; nor do we".[29] While robots that shed tears have been built too,[30] for anyone who considers apologies as expressives that report an emotional state, machine apologies are impossible.

But while this is one way of thinking about apologies, it is not the only one. As noted above, apologies are used extensively as a managerial tool, and a considerable amount of the literature on the trust-repairing effect of apologies issued by, or on behalf of, companies. Apologies also play an important role in post-conflict societies, and have been instrumental in quasi-judicial procedures such as the South Africa Truth and Reconciliation commission. When Tony Blair apologised for Britain's role in the slave trade, he will not have felt personal remorse.

It is true that the lack of remorse, or personal responsibility, is often seen as cheapening the currency of apologies and potentially manipulative.[31] But if these apologies are manipulative, then it is a manipulation where we are all willing and informed participants – nobody thinks that really, a spokesperson for a government or a company "feels remorseful" when saying what their job requires them to say, and despite this knowledge, the "healing effect" is real and measurable.[32] Furthermore, not only are these "public apologies" intelligible to us, we still distinguish successful from unsuccessful apologies, legitimate from illegitimate ones.

This allows us to identify criteria that are needed so that the apology restores trust, criteria that can be different from those we use when humans apologise for their own actions.[33] For this reason, we will for the rest of this chapter talk of and contrast two types of apologies. One is the "remorse

[28] See e.g. Kate Rossmanith, 'Affect and the Judicial Assessment of Offenders: Feeling and Judging Remorse.' (2015) 21 Body & Society 67; Margreet Luth-Morgan, 'Sincere Apologies: The Importance of the Offender's Guilt Feelings' (2017) 46 Neth. J. Legal. Phil. 121.

[29] STATE v. THORNTON (2002) Nos.99-376-C.A., 98-263-C.A.

[30] Akiko Yasuhara, Takuma Takehara, 'Robots with Tears can Convey Enhanced Sadness and Elicit Support Intentions. (2023) 10 Frontiers in Robotics and AI 1121624.

[31] So e.g. Lee Taft, 'Apology Subverted: The Commodification of Apology.' (1999) 109 *Yale lJ* 1135.

[32] See e.g. Michael R Marrus, 'Official Apologies and the Quest for Historical Justice' (2007) 6 Journal of Human Rights 75.

[33] Taenyun Kim, Hayeon Song, 'How should Intelligent Agents Apologize to Restore Trust? Interaction Effects between Anthropomorphism and Apology Attribution on Trust Repair' (2021) 61 Telematics and Informatics 101595.

expression" (RE) apology that we use for personal wrongdoings between people. The other is the "public apology" (PA). While the way these are expressed in language is in parts similar, and in particular shares expressions such as "sorry" or "I apologise", they do have their own distinctive logic.

A related objection is that a sincere apology requires that it is given voluntarily. Blair may not have felt personal remorse, but the decision to apologise on behalf of the UK came with political risk that he was willing to take. By contrast, the Australian Prime Minister Howard refused to apologise on behalf of the Australian government.[34] In each case, the ethical salience, and the effect on trust in their leadership, might reside in the fact that they could have done otherwise. The Yale robot did not have this choice, and maybe this lack of freedom undermines, or should undermine, any assessment of its sincerity. And indeed, we find that the more schematic and "enforced" a robot apology is (think as an extreme example of 404 error messages), the more its sincerity is doubted.[35] But in law, apologies can also be ordered by a court as a civil remedy.[36] The historical precursor of John Wayne's bon mot dates back to 1869 when the *New York Tribune* criticised *The Times* for reversing an editorial position without openly admitting the change:

> "It never apologizes, never retracts, never allows its readers to remember that it is eating its own words"[37]

Depending on jurisdiction, today *The Times* may find itself ordered by a court to print an apology[38], or at least face more severe sanctions by its regulator for violations of the Editors' Code if no apology is forthcoming.[39]

[34] Mary R. Power, 'Reconciliation, Restoration and Guilt: The Politics of Apologies', (2000) 95 Media International Australia 191.

[35] Xingyu Wang, Yoo Hee Hwang, Priyanko Guchait, 'When Robot (vs. Human) Employees Say "Sorry" Following Service Failure.', 24 (2023) International Journal of Hospitality & Tourism Administration, 540.

[36] Brent T. White, 'Say you're Sorry: Court-Ordered Apologies as a Civil Rights Remedy' (2005) 91 Cornell L. Rev 1261.

[37] 1869 March 9, New-York Tribune, Foreign News: The Rejection of the Alabama Convention, Quote Page 1, Column 4, New York, New York. From https://quoteinvestigator.com/2023/01/20/howl/#320b2489-64e7-486e-afb0-e18cd36f7eba.

[38] Wannes Vandenbussche, 'Rethinking Non-Pecuniary Remedies for Defamation: The Case for Court-Ordered Apologies', (2020) 9 J. Int'l Media & Ent. L. 109.

[39] http://www.editorscode.org.uk/downloads/codebook/codebook-clause-1.pdf.

While it is true that the wisdom of court-ordered apologies is controversial[40], they are no doubt intelligible as apologies.

How do (forced or freely given) public apologies function without an internal feeling of remorse? To answer this question, we have to ask why apologies can restore trust in the first place.

One way to account for RE apologies as *rational* trust repair is that they give us good reasons to believe that the same harm won't occur in the future. 'Moral emotions" such as remorse matter for both ethics and psychology because the express our ability to self-reflect[41]. With the ability for self-reflection comes the ability to understand where we went wrong – and with that also the ability to correct our behaviour next time round. Apologies externalise this internal mode of reflection. As Tavuchis puts it, an apology is a performative utterance that in the case of RE converts the remorse of the offender from "a private condition into public communion".[42] Remorse, especially remorse that reaches the level of pain, is then the motivating factor that allows us to conclude that the apologiser will act on their insight. The first apology I gave at the beginning of this paper failed because it was immediately followed by a pragmatic retraction, my announcement that I would keep sinning For Tavuchis, the promise of change is so inextricable intertwined with the expression of remorse that it does not even need saying, it is always implied.[43]

PA and RE share this external form of "public communion". What is missing in many forms of PA is the internal, motivating factor for change. How can it be replaced? If the role of remorse is as warrant for a future change in behaviour, then an externally enforceable promise of change can take its place for the purpose of trust repair. Here the law can come into play. Change when making a RA is an implied consequence that "may" not need stating explicitly, because our folk psychology tells us that the pain of remorse will lead to change in behaviour[44]. In a PA, this commitment to change becomes part of the felicity conditions of a successful apology

[40] Nick Smith, 'Against Court-Ordered Apologies', (2013) 16 New Criminal Law Review 49.

[41] Jerome Kroll, and Elizabeth Egan, 'Psychiatry, Moral Worry, and the Moral Emotions' (2004) 10 Journal of Psychiatric Practice 352. For a philosophical discussion see Benjamin Vilhauer, 'Kantian Remorse with and without Self-Retribution' (2022) 27 Kantian Review 421.

[42] Tavuchis, (23) 64.

[43] Ibid. 23.

[44] Many theorists of apologies suggest that also an effective RE will normally require an offer of reparation and/or promise of change. See e.g. Aviva Orenstein, 'Apology

that need to be communicated and stated explicitly. Apologies then do not so much report an internal state of regret, rather they report a line of reasoning where the offender

1. takes responsibility, which includes a causal account of the actions and conditions that led to the harm
2. states the steps that will be taken to prevent the same mistake happening again
3. possibly makes an offer of "making good" – compensation that is commensurate to the harm inflicted and the degree of responsibility

1) turns the apology into the exact mirror image of the account of "explanation" that we discussed above. Just as an explanation is not an account of the inner processes that led someone to reach the right result, but a publicly verifiable account of a valid chain of reasoning, so is an apology often not an account of the inner processes that led to a mistake (that would be "making excuses") but a publicly auditable account of what caused the harm. Because of this symmetry, I argue that for regulatory purposes, we should consider this form of apology as an appropriate way to meet explainability requirements, even if it does not disclose the inner working of the AI.

We can now also express more clearly the objections against the Yale robot. The problem is not that it deceives its audience by claiming to have an internal state that machines do not possess – its machine nature is too obvious for this. Rather, by using the verbal form, or logic, of an RA apology, it deceives us in inferring also conditions 1-3, that is we falsely infer that:

1) the reason the robot states for its failure is the causally relevant reason (in the example: processor not fast enough)
2) that the same issue will not happen again, that there will be change.

In the experiment the robot's "failure" was externally enforced, not the result of an unsuitable processor being used for the task, it was a "placebo explanation" that unfortunately can be as efficient in restoring trust as real explanations.[45] Because the causal explanation is already false, there is no pathway from a recognition of responsibility to an effective change in future

Excepted: Incorporating a Feminist Analysis into Evidence Policy Where You Would Least Expect It' (1999) 239 Southwestern U. Law. *Rev* 221.

[45] Malin Eiband and others 'The Impact of Placebic Explanations on Trust in Intelligent Systems.' In *Extended abstracts of the 2019 CHI conference on human factors in computing systems*, (ACM 2019) 1 https://dl.acm.org/doi/10.1145/3290607.3312787.

behaviour. Furthermore, the robot does not have the capacity to effect the change that its own words indicate. It cannot for instance upgrade itself, or refuse to play next time it encounters a scenario that requires speedy responses.

A PA apology that does not just restore trust, but trustworthiness, therefore requires at the bare minimum a correct causal account of the contribution that an action or omission had for the harm, and an enforceable or auditable promise of future change. Not

> "Sorry, I sometimes run out of memory and can't process things fast enough".
> But
> "I'm sorry, I was not fast enough in move 3, I'm going to download an upgrade before the next game, and I'll also pay the participation fee for the next round of gaming".

This structure of an apology also mirrors definitions found in some legal systems. The Apology Scotland Act (2000) for instance defines as valid apology for the purpose of the act

> "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."

If we read the expression of regret as acceptance of (causal) responsibility, then the "undertaking to look into the circumstances [...]" equates to a causal explanation of why the harm occurred, while the prevention element is forward looking and gives reason to restore trust.

The purpose of this Act is to encourage apologies, which are often what the victim prefers over other remedies, but which are often actively discouraged by an institution's or corporation's lawyers.[46] Apologies, so their reasoning, can be constructed as admissions of legal liability, even though the legal requirements may be much more exacting than a personal feeling of responsibility. [47] Apologies that conform with the structure prescribed

[46] Elizabeth Latif, 'Apologetic Justice: Evaluating Apologies Tailored Towards Legal Solutions', (2001) 81 Boston University Law Review 28.

[47] See Jennifer K Robbennolt, 'Apologies and Legal Settlement: An Empirical Examination' (2003) 102 Michigan Law Review 460.

by the Act are privileged for the purpose of litigation, that is they can't be adduced as evidence by a complainer if they decide to sue for damages.[48] This does not mean that receiving an apology bars them from brining an action for damages, only that they need to find evidence other than the apology itself.

This type of liability shield should be attractive to anyone who contemplates allowing robots to apologise on their behalf, as it mitigates the risk of apologies that turn out to have been premature, or in some other way not merited. At the same time, because apologies that are valid for the purpose of the Apology Act are also the mirror image of explanations, they help complying with transparency requirements under instruments such as the EU AI Act.

E. Robot-Love means never having to print 1001001001010

The second ConTrust working paper stated:

"Normatively justified trust relations in situations of conflict come about and persist when the right to justification (in a broad sense) is in place despite and in light of conflict."[49]

In this chapter, I tried to argue that if XAI were to take this insight at heart (as it should), we need in addition to "cooperative AI explanations" of the type WMR developed also "conflict-centric" XAI that operates after norms or reasonable expectations were violated by an AI.

Having identified apologies as the type of speech act that meets the requirements for a trust-restoring, conflict-centric "explanation", the question became whether AIs are in principle capable of apologising, arguing that most if not all currently developed "apologetic" robots are undermining rather than enhancing justified trust. Drawing from ideas across a range of legal disciplines, I argued that we must distinguish two different types of apologies. One relies on internal mental states as a guarantor for change, the other on making actionable promises in a public forum. The former can only be performed successfully by humans, the latter also by companies, states or other abstract entities and the people who speak on their behalf. The reason that the Yale robot (and others like it) are ethically dubious is

[48] Prue Vines, 'Apologies and Civil Liability in the UK: a view from elsewhere. (2008) 12 Edinburgh Law Review 200.

[49] Forst (1) 8.

that they are not the right type of agent to make the first type of apology, and on the other hand they are not using the correct format for participating in the second type of apology game.

For the normative issues, we concluded that to the extent that robots apologise, they should always only use type 2 (PA) apologies. If they use PA apologies, then this is a valid way to establish the type of explainability that legal instruments such as the AI Act mandates. To encourage building the capacity of type 2 apologies into AIs, we should consider for those jurisdictions that give litigation privileges to human-authored apologies to extend them to robot-issued apologies.