



Against self-incrimination in the AI era: Inferred emotions seeking for protection

By Paul De Hert* and Georgios Bouchagiar*

* Paul De Hert, Full Professor at Vrije Universiteit Brussel, Associated Professor at Tilburg University, Honorary Director of the Brussels Privacy Hub, Paul.De.Hert@vub.be

* Georgios Bouchagiar, Postdoctoral Researcher at Vrije Universiteit Brussel, Georgios.Bouchagiar@vub.be

The Brussels Privacy Hub publications are intended to circulate research in progress for comment and discussion. Available at <https://brusselsprivacyhub.com/>. Reproduction and translation for non-commercial purposes are authorised, provided the source is acknowledged.

DISCLAIMER

The opinions expressed in this paper are those of the author/s.

Contents

1	Introduction: Justice Brandeis' prediction on advances in psychic and related sciences	4
2	Defining emotions and their inferring and the need to look at it through the lens of self-incrimination	5
3	Deficiencies in the current EU legal framework on data and AI	6
4	The human rights perspective on self-incrimination in Europe.....	7
5	Three considerations that explain why inferred emotions <i>may</i> fall outside the protective scope of the right against self-incrimination	8
6	The argument for and steps towards protecting emotions against machine-inferring	10
7	Conclusion: returning to Justice Brandeis' wisdom	11

Abstract

Emotion-inferring via AI is defended for its promising benefits for society, such as effectiveness in combatting crime. However, it largely seems to escape protection granted under the right against self-incrimination. This post proposes the recognition of 'inferred emotions' as a new category of protected communications, with a view to effectively regulating, and even banning, such AI-uses.

1 Introduction: Justice Brandeis' prediction on advances in psychic and related sciences

Olmstead is a famous 1928 US Supreme Court case on telephone tapping by federal officers. The defendants objected to admission of evidence obtained by wiretapping on the ground that wiretapping constituted an unreasonable search and seizure in violation of the [Fourth Amendment](#) and that the use as evidence of the conversations overheard compelled the defendants to be witnesses against themselves in violation of the [Fifth Amendment](#). Justice Brandeis, in his dissent, argued forcefully that the two constitutional rights had to be taken seriously, in light of new technological developments. When the Fourth and the Fifth Amendments were adopted, '*the form that evil had theretofore taken*' had been necessarily simple ([Olmstead](#), para 473). '*Force and violence were then the only means known to man by which a Government could directly effect self-incrimination*' ([Olmstead](#), para 473). Both rights offered solid protection against force and violence, but subtler and more far-reaching means of invading privacy had become available to the Government in the meantime. Brandeis argued for interpretative action: '[t]he progress of science in furnishing the Government with means of espionage is not likely to stop with wiretapping (...) Advances in the psychic and related sciences may bring means of exploring unexpressed beliefs, thoughts and emotions' ([Olmstead](#), para 474).

Today, Brandeis' prediction on such advances has become a reality: emotions can be explored via AI. With both [Europe](#) and the [US](#) experimenting on emotion-inferring, we may soon witness law enforcement using technologies to foretell our innermost data on the basis of our facial micro-expressions. And, if we risk being mistreated on the basis of voluntary or involuntary manifestations of our body, we may find ourselves in Orwellian dystopias, either having to hide or fake how we feel to protect ourselves from intrusion into the conscious or, worse, having to tolerate interference with the unconscious.

This post focuses on emotion-inferring via AI and the right against self-incrimination. After some general remarks on emotions and their inferring, as well as European laws on personal data and AI that may become relevant/applicable, we move on to the gist: the implications of AI-based emotion-inferring for the right against self-incrimination. Our post analyses the European ([ECtHR/CJEU](#)) approach, with a view to exploring whether and the extent to which the existing account of this right is ill-equipped to address challenges posed by such AI-uses. Finding that emotion-inferring may

remain unprotected, it suggests the recognition of ‘inferred emotions’ as a new category of protected communications, with the objective of effectively banning or otherwise regulating AI-uses intended for emotion-inferring.

2 Defining emotions and their inferring and the need to look at it through the lens of self-incrimination

‘Emotions’ are mental states. They are experiences or responses to events, which may be automatic, but not always unconscious, and which can involve coordinated changes across physical sensations, thoughts, action tendencies and/or behaviours. This definition, in line with psychology, neuroscience and other perspectives, sees emotions as states involving the ‘conscious’ and the ‘unconscious’, as well as the ‘physical’ and the ‘mental’ (see, among others: [Schacter, Gilbert and Wegner](#); [Ekman and Davidson](#); [Panksepp](#); [Feldman-Barrett, Lewis and Haviland-Jones](#); [Engelen and Mennella](#)).

Thereafter, ‘emotion-inferring’ refers to the processing of observable data (like facial expressions) with the goal of interpreting these data and, ultimately, attributing mental states (emotions) to them. While observable data can be ‘physical’, the interpretation step crosses from physical manifestation to ‘mental’ state attribution.

That emotion-inferring can be performed via AI means that the data processing can be conducted in a black-box-fashion – an opaque process that human experts may not fully comprehend. Despite lack of transparency, same as many (if not all) AI-implementations, emotion-inferring has been widely adopted in the private and public sector for numerous purposes (e.g., [predicting behaviour of job candidates](#), [identifying fraudulent conduct of customers](#), [checking students/pupils’ attentiveness](#), [detecting lie](#) or [foreseeing violence](#)).

To its advocates, emotion-inferring can promise benefits for society, such as effectiveness in combatting crime (e.g., [managing border-crossing or screening public spaces](#)). But, to commentators, these AI-implementations [could raise serious concerns on human rights/freedoms](#). While literature has paid much attention to privacy, freedom of expression or undue discrimination (e.g., [EPRS](#); [Uguz](#); [Levantino](#); [FRA](#)), the right against self-incrimination is mostly overlooked. This may be surprising, in light of the above AI/emotion-practices, as well as the age-old belief that the protection of emotions and other mental states is the *raison d’être* of the right against self-incrimination ([Dann](#), p. 611).

3 Deficiencies in the current EU legal framework on data and AI

Before we look at the right against self-incrimination in the context of AI-based emotion-inferring, two remarks need be made on data protection- and AI-related laws that may, in view of the above approach to emotion-inferring, become applicable.

EU data protection laws – the [GDPR](#) and the [LED](#) – set out concrete processing principles regarding personal data and imposed rigorous duties on data controllers. Yet, applicability of such laws is doubtful, in light of lack of consensus, first, on the definition of ‘emotion’ ([Posner](#), p. 1) and, second, on qualifying emotions as personal data ([Duchonová](#), p. 25). In any event, even if it were accepted that emotions can qualify as personal data, neither legislative act contains specific rules, be it enabling or prohibitive, on emotion-inferring or AI. Both are technologically neutral and abstract ([De Hert and Bouchagiar](#), p. 206) and, hence, do not expressly address particularities of AI-implementations aimed at emotion-inferring.

The [AI Act](#) establishes a common regulatory framework for AI systems within the EU. Contrary to data protection laws discussed *above*, this Act is the opposite of technological neutrality: it is targeted at AI systems that meet concrete minimums (Article 3(1); [Guidelines on the definition of an AI system](#)) – notably, risking creating loopholes in its substantive application-scope, in view of future AI-developments/advancements (that might go beyond the Act’s definition of an AI system). Relevant here is the attempt in the Act to regulate emotion-inferring via AI. It accepts that the processing of biometrics can allow for emotion-recognition (recital 14); and that AI systems can infer/identify emotions on the basis of biometrics (Article 3(39)). However, it does not define ‘emotions’. Instead, it only gives examples of emotions/intentions (e.g., happiness or anger); and it clarifies that it does not cover physical states (e.g., fatigue) or the identification of ‘*readily apparent expressions, gestures or movements*’ (e.g., facial expressions), ‘*unless they are used for identifying or inferring emotions*’ (recital 18); meaning that facial expressions can be covered, when used for emotion-inferring ([Guidelines on prohibited AI practices](#), para 249).

The AI Act acknowledges technological and other shortcomings of emotion inferring/identifying, including ‘*limited reliability, the lack of specificity and the limited generalisability*’, possibly resulting in undue discrimination or other interferences with human rights/freedoms (recital 44). For this reason, it prohibits AI-uses aimed at emotion-inferring in the areas of education and work (save for safety and medical reasons; Article 5(1)(f)). It also classifies AI-uses aimed at emotion-recognition as high-risk in other areas (Article 6(2); Annex III, para 1, lit c). This means that those designing/deploying AI systems must comply with certain obligations and requirements (Articles 8-27). Still, such obligations/requirements (especially for emotion-inferring) appear limited, accompanied by exceptions favouring law enforcement. A prime example is

transparency and related obligations (e.g., to inform an individual that she/he is exposed to emotion recognition via AI). Such duties, in principle, do not apply to emotion-recognition for criminal investigations (Article 50(3)).

Carefully, we can conclude that the EU data protection laws lack clear rules targeted at AI-based emotion-inferring; while the AI Act fails to define ‘emotions’ and appears rather permissive, allowing for exceptional uses by law enforcement.

4 The human rights perspective on self-incrimination in Europe

Let us move on to human rights law as a framework and complement to the laws discussed *above*. We recall the respective roles of the [Council of Europe](#) (CoE) and the [EU](#). The former includes a larger number of addressees and its key human-rights-instrument is the [ECHR](#), interpreted by the [ECtHR](#). The latter has a smaller audience, its primary tool on human-rights-protection is the [Charter](#) and the [CJEU](#) supervises compliance with EU law. In general, the EU legal framework has been inspired by and has heavily relied upon the CoE’s legal regime ([Craig and de Búrca](#); [Schütze](#)). Neither the ECHR nor the Charter include an explicit clause against self-incrimination.

The ECtHR has recognised the existence of such a right within Article 6 ECHR (on fair trials). This right protects suspects against improper compulsion ([Ibrahim](#), paras 266-267). [Saunders](#) (paras 68-69) specified that this right requires that the prosecution prove their case without reliance upon evidential materials that were gathered via ‘*coercion or oppression in defiance of the will of the accused*’; this protection, however, does not cover evidence that exists independently of the suspect’s will – e.g., documents obtained through warrants, breath/blood for DNA-testing or other ‘real’ evidence, compared to forced ‘statements’.

Although the ECtHR seems to differentiate between ‘statements’ and ‘real evidence’, there have been cases, where the distinguishing line appears blurred. Of relevance is [Jalloh](#), where the ECtHR found a violation of the right not to self-incriminate: the forcible administration of emetics to obtain evidence from the applicant impaired the very essence of the protection against self-incrimination (see in particular paras 97-102).

The *Jalloh* judgment seems to assume that the taking of real evidence may require that suspects passively undergo an insignificant interference with bodily integrity (e.g., DNA-sample-taking) or may demand active participation when this active involvement is aimed at getting materials created by the physical *modus operandi* of the body (e.g., urine). In this case, forced emetics, demanding the use of a tube through the nose to provoke a reaction in the body ([Jalloh](#), para 114), thus, resulting in a heavier interference, were not tolerated. Amongst the key factors to find a violation of the right against self-incrimination are elements,

such as the nature/degree of compulsion (seriously interfering with physical/mental integrity); the weight of the public interest (in securing conviction for drug-selling on a rather small scale) and the punishment of the crime (suspended prison sentence/probation), hardly justifying the interference; the existence of procedural safeguards (legal basis protecting against health-risks); and the role evidence played in decision-making (drugs collected via forced emetics played a determinative role in conviction) - ([Jalloh](#), paras 101 and 118-121).

The EU's approach to fair trials, in general, and the right against self-incrimination, in particular, has been inspired by the CoE's legal regime ([Explanations relating to the CFR](#), Articles 47-48). Notably, the EU follows a similar approach to this right (e.g., Article 7(3) [Directive \(EU\) 2016/343](#)), that the CJEU has addressed as an internationally accepted standard that '*lies at the heart of the notion of a fair trial*' ([Consob](#), paras 37-38, interpreting Articles 47 and 48 [CFR](#) in view of ECtHR's case law on Article 6 [ECHR](#)).

5 Three considerations that explain why inferred emotions *may* fall outside the protective scope of the right against self-incrimination

Self-incrimination-protection is about respecting the accused's will to remain silent. Europe seems to protect statements that are 'forced' and 'incriminating' (e.g., forcible evidence-obtaining leading to criminal prosecution). Emotion-inferring via AI *could* fulfil these minimums: e.g., involuntary processing of biometrics with a view to inferring frustration that may then be used to establish criminal liability (or, contrary, voluntary biometric processing that is nevertheless aimed at inferring emotions against the will of the suspect and, ultimately, establishing criminal liability).

We write 'could'. Several considerations explain our prudence.

First consideration is about the binary *real evidence versus statements*. While communications revealing the content of the mind (statements) are protected, physical/real materials are not. The problem with emotion-inferring is that it does not clearly fit into either category: its input is physical (e.g., facial expression, coming as a manifestation of a specific 'true' emotion), but its output is a claim on emotions (like fear, 'inferred' from that facial expression, but not necessarily identifying with the 'true' emotion). Relevant here is to understand that, unlike purely physical evidence (such as DNA that is 'what it is', our genetic code), facial or other observable data do not directly equal to emotions but the result of interpretation.

Second consideration is about the will-criterion. The European approach would ask whether inferred emotions can exist independently of the suspect's will ([Saunders](#), paras 68-69). Admittedly, one cannot always 'will' her/himself to or not to feel emotions, especially unconscious emotions. For instance, one may

feel angry and have the tendency to lash out, but instead deliberately chooses to act calmly (e.g., by hiding/faking her/his facial expressions); but this may not be the case with unconscious reactions, like surprise arising automatically prior to conscious thoughts (that would otherwise enable someone to hide/fake her/his emotions).

A *third consideration* regards the criterion of non-invasiveness and a possible parallel between DNA/urine, on the one hand, and emotions, on the other. Can one argue that inferred emotions are created by the normal operation of the body, involving no significant interference with physical/mental integrity (like urine/DNA-taking; [Jalloh](#), para 114)? It is possible to hold that facial expressions or other observable data can be linked to the physical operation of the body and may be processed (e.g., by a camera) without seriously affecting bodily integrity. Under the European legal regime ([Jalloh](#), paras 118-121), inferred emotions could escape protection, if and to the extent that: the revelation of emotions against the suspect's will were regarded as a minor interference; the interference were justified by the public interest; procedural safeguards were in place; or inferred emotions played a non-determinative role in the decision-making process.

But can the DNA/urine-analogy accurately and fully capture emotion-inferring's particularities? Although emotion-inferring can (same as urine/DNA-taking) be physically non-invasive, it may reveal mental content against the suspect's will.

In light of the above considerations, inferred emotions may fall outside the protective scope of the right against self-incrimination, as framed under the European legal regime. And, even if protection were granted, this would, in the absence of concrete legal provisions on emotion-inferring, come *ex post*; e.g., in view of and by analogy with related case law, via a judicial decision finding violation of the right against self-incrimination; this, after the materialisation of the risk –and the harm.

This 'by analogy' and 'on a case-by-case basis' protection could create legal uncertainty and fail to adequately protect against intrusive mindreading methods targeted at the unconscious. In fact, previous experience on brain imaging/reading ([Fox](#); [Escobar Veas](#)) or other methods aimed at extracting mental states and content from the body ([Ligthart](#); [Uviller](#)) has demonstrated how contemporary accounts of the right against self-incrimination may remain ill-equipped to sufficiently protect the suspect. In view of such legal uncertainty and failures, in what follows, we propose the recognition of 'inferred emotions', as a new category of protected communications (alongside the existing categories of 'real evidence' and 'statements').

6 The argument for and steps towards protecting emotions against machine-inferring

Emotion-inferring can create forced disclosure of internal states that the right against self-incrimination was designed to prevent ([Doe](#), para 211; [Dann](#), p. 611). Yet, it does so via a tech-route (the interpretation step) that may sidestep traditional protections. We propose to recognise explicitly ‘inferred emotions’ as a new category of protected communications and bring it under the protective scope of self-incrimination to protect against the developing use of emotion-inferring performed by AI. Several steps are needed to achieve this goal.

The first step is definitional: ‘inferred emotions’ should be clearly defined. A definition, similar to that given in the beginning of this post, could to some extent reflect emotion-inferring’s key functions: processing of observable data that can often refer to involuntary/uncontrollable manifestations; interpretation of these data; and emotion attribution. This first definitional step could assist regulators in overcoming legal uncertainty of case law (adjudicating on a case-by-case basis, thus, only by analogy applicable to ‘inferred emotions’); but also in addressing weaknesses of existing AI-related legal instruments to clearly define emotion-inferring (e.g., [AI Act](#), recital 18, referring to examples of emotions, instead of defining them).

The second step is normative and consists in assessing whether/how emotion-inferring can threaten key values underlying the right against self-incrimination, e.g.: mental integrity (e.g., accessing/revealing mental states without consent); dignity (e.g., rendering humans readable objects); or fairness of trials (e.g., instead of having the state proving guilt without the defence’s forced cooperation, the defendant may be compelled to participate in her/his own conviction via uncontrollable emotional responses). These threats appear increased for unconscious states of mind, often uncontrolled/unknown by/to the suspect. Moreover, for both conscious/unconscious emotions, these risks can be augmented, in light of emotion-inferring’s unique unreliability-concerns, that have already been recognised by law (e.g., [AI Act](#), recital 44), e.g.: context-dependence (the same expression may mean different things); cultural/individual-variability (emotion-expressions may vary across cultures or individuals); and error-rates.

A third step would be to complement the premature, incomplete attempt in the AI Act to regulate, and where necessary, ban AI-uses that are considered illegitimate under the light of the normative exercise *above*. An avenue for concrete reform would be an outright ban on unconscious emotion-inferring – that can escape protection granted under the European approach. In addition, presumptive inadmissibility could be introduced for all types of (conscious/unconscious) emotion-inferring – e.g., unless the suspect consents to

its use and/or the state shows that there can be a compelling public interest *and* that emotion-inferring is the least intrusive means.

In the latter cases (e.g., where emotion-inferring is used upon consent), there could be stringent scrutiny on the basis of reliability-requirements, such as the minimums known from US case law on admissibility ([Frye](#), [Kelly](#) or [Daubert](#)) (see, in this regard, [De Hert and Bouchagiar](#), pp. 18-19). For example, [Daubert](#)'s requirements of adequate testing/review, no/low error-rates or acceptance by the relevant scientific community might hardly be met by contemporary AI-implementations.

Furthermore, and again for all types of (conscious/unconscious) emotion-inferring, contextual/procedural safeguards could be added to (and support) existing ones (known from settled case law, like [Jalloh](#)), e.g.: emotion-inferring should in *no way* be used as a decisive/determinative factor, but only in conjunction with other materials that can independently support the judicial or other decision; judges or other public actors should be informed on tech-limitations; the defence should be given the chance to challenge emotion-inferring via its own expert; there should be prior (e.g., warrant-like) authorisation for using such tools and clear notification of the suspect; or the suspect should enjoy the right to refuse subjection to these tools (without suffering any detriment for refusal).

In this way, the proposed scheme, introducing a new category of protected communications, could assist in overcoming legal uncertainties and failures (e.g., by providing a uniform and standardised approach to emotions and their inferring), better comprehending the risks stemming from AI-based emotion-inferring-implementations vis-à-vis the right against self-incrimination, as well as regulating emotion-inferring in an *ex ante* and targeted fashion, taking into due account the particularities of AI/emotion-inferring, but also building on and supplementing well-established safeguards known from settled case law.

7 Conclusion: returning to Justice Brandeis' wisdom

Let us return to Justice Brandeis in *Olmstead*. Our human criminal law-related rights saw the light when the form that evil had theretofore taken (force and evil) had been necessarily simple. Sticking to the old criteria for applying and recognising the protection for self-incrimination is not seeing the more 'subtler and more far-reaching means of invading privacy available to the Government' and 'the progress in science'. If we want to continue with a legal system that only protects against 'breaking into a house and opening boxes and drawers' and ignore the 'invisible and intangible', we sacrifice basic values and let governments play with our freedom by '*exploring unexpressed beliefs, thoughts and emotions*' ([Olmstead](#), paras 474-475). Brandeis advocated for giving effect to the principles underlying the criminal law-related human rights (such as the

Fourth and the Fifth Amendment) and refused ‘to place an unduly literal construction upon it’ ([Olmstead](#), para 476).

This post assessed the European legal framework on the right against self-incrimination *vis-à-vis* AI-based emotion-inferring. Finding that emotion-inferring can escape protection, it recommended the recognition of ‘inferred emotions’ as a new category of protected communications, meriting special protection and attention. The proposed approach could assist regulators in clearly defining emotion-inferring, detecting its key challenges and, ultimately, effectively banning or otherwise *ex ante* regulating AI-uses aimed at emotion-inferring.

To conclude, if it is true that the protection of emotions and other mental states is at the heart of human-rights-related laws ([Olmstead](#), para 478); and, if it is accepted that this protection can today be under an unprecedented attack by AI/emotion-experiments that are increasingly being conducted around the globe (e.g., the [EU](#), the [UK](#), the [US](#) or [China](#)), by AI/emotion-practices that are engaged in by more and more public actors (e.g., the [police](#), [army departments](#) or [councils](#)) or by ever-growing AI/emotion-surveillance that appears to shift from the mass and the collective toward the personal, [each one of us](#); then, there is an acute need for stringent regulatory intervention against self-incrimination, with a view to protecting ourselves from being unfairly judged on the basis of our appearances, rather than our conscious actions.