Social affective inferences in the era of AI-filters: towards the Bayesian reshaping of human sociality?

Nadia Guerouaou[a,b,c], Guillaume Vaiva[c], JJ Aucouturier[a]

[a]*FEMTO-ST Institute, Besançon, France*
[b]*STMS, Paris, France*
[c]*Lille Neuroscience & Cognition, Lille, France*

Abstract

In a world where our social interactions are increasingly mediated by digital technologies, the possibility to use AI *filters* to artificially control how we appear and sound to one another is quickly becoming a commodity. Here, we propose the framework of Bayesian inference as a way to question how "externalizing emotions" in such a way would affect our social-cognitive functions, both during such interactions and in society as a whole.

The expression of emotions in human and its AI-filtered digital self

In our daily interactions, how we speak and appear to others conveys a rich para-linguistic stream of mental states and attitudes that, often, are as predictive of interpersonal consequences as our words themselves [1]. These may include facial and vocal cues of e.g. emotions, social attitudes such as warmth or dominance, or epistemic information about the certainty of a spoken fact. The theory of predictive inference [2] posits that, as other judgements about the world, the meanings we ascribe to such social expressions are based on an internal generative model formed of beliefs learned from experiencing similar expressions in other contexts. In short, the varied expressions we encounter in our social environment today shape the perceptual inferences that underpin our future social behaviours, from maintaining successful relationships to scheming for political power.

Email address: n.guerouaou@gmail.com (Nadia Guerouaou)

While facial and vocal expressions have long been considered natural and honest cues to infering ones' emotions and attitudes [3], new and rapid technological developments are fundamentally altering their informational nature. In just a few post-covid years, video-communication interfaces have become an integral part of our personal and professional lives and, with the ongoing progress of artificial intelligence (AI), such communication is increasingly mediated by augmentative technologies that allow to control the way we appear and sound to others. This is recently illustrated with e.g. the surge of beauty or smile filters on platforms like Instagram or TikTok [4], or voice enhancement in gaming communities on Discord. Addressing the mediating effects of AI on self-presentation and social relations therefore seems essential in a world where boundaries between the digital and physical are increasingly porous. In the line of Clark and Chalmer's extended mind theory [5], it is now our socio-affective processes that are "externalized" in our technologies: with the almost ineluctable adoption of such technologies, we are approaching a situation in which the appearance of our social behaviours finds itself under our direct and independent technological control. This situation of algorithmically releasing the biological constraints of facial and vocal expressions in our social interactions, is in our view unprecedented in the history of human societies. Here, we propose the framework of Bayesian predictive inference as a way to think and question how AI filters of face and voice might come to affect our social-cognitive processes and, over time, human sociability in general.

A Bayesian account of social-affective evaluation of other's

The theory of predictive inference suggests that the brain infers the most likely causes of its sensory inputs by minimising the difference with signals predicted on the basis of previously-learned generative models [2]. Such predictions are typically described in the mathematical framework of Bayes' law, per which the a-posteriori probability of hypothesis $\mathcal{H}$ given a certain observation $\mathcal{O}$ depends on the likelihood $p(\mathcal{O}|\mathcal{H})$ of similar observations in our past encounters with $\mathcal{H}$, and on our belief in the a-priori plausibility of $\mathcal{H}$ and $\mathcal{O}$:

$$p(\mathcal{H}|O) = p(O|\mathcal{H}).p(\mathcal{H})/p(\mathcal{O}) \tag{1}$$

In social interactions, Bayes' equation can be used to explain how we infer mental states from facial and vocal cues received from others. Take, say, an encounter with our friend Paula in a Parisian café . Paula is smiling (our observation $\mathcal{O}$), and we may want to decide whether that means she's happy to see us (our hypothesis $\mathcal{H}$). Bayesian inference says this evaluation is based, first, on the likelihood $p(\mathcal{O}|\mathcal{H})$ to see Paula smile in that particular way when she is happy (as opposed to how she'd smile in domination or sarcasm). This quantity is then modulated by the prior probability of hypothesis $p(\mathcal{H})$ regardless of observation (e.g. Paula is our friend, so the probability that she'll be happy to see us is high) and that of the observation $p(\mathcal{O})$ regardless of context (which may depend e.g. on the personality of the observee and the display norms of a particular culture).

Critically, all three factors are estimated from probability distributions learned by the observer from previous experience. Therefore, increased exposure to novel and rapidly changing forms of affects display in AI-mediated communication may have profound short-term and long-term effects on how we judge mental states online and, quite probably, offline.

How AI face and voice filters could reprogram social inferences

Let's take our interaction with Paula online, where she has the option to use tools such as "smile filters" that, without our knowing, allow her to control how she looks and sounds to us. Bayes' equation provides a framework in which to think of how such interventions could affect social inferences, in the short, middle and long term.

The first, and most immediate, effect of infering mental states in a AI-filtered world is committing what we might call "IRL-centric" errors: by modifying observations $\mathcal{O}$, AI filters modify the probability we assign to them using probability distribution $p(\mathcal{O}/\mathcal{H})$, which we learned from "real-life" situations (i.e. offine, or online without filters) in which smiles were used more genuinely. The resulting inference errors may lead one to ponder e.g. why everyone on zoom appears so friendly - or why our Paula is so visibly happy to see us today. Such effects are insidious, both because of the realness of recent AI transformations and because inference processes often being non analytical. It is therefore unclear whether interventions that simply label these expressions as "technologically-transformed" would suffice to compensate inference errors. Besides, these effects may be diverge from what these tools are supposedly designed for, and thus be hard to predict: for instance, generative algorithms trained on seemingly neutral datasets have the unexpected effect of generating faces that are more trustworthy than average [6]. In our view, understanding how exactly AI-filters change sensory observations and how observers evaluate their likelihood will become critical subjects for psychological behaviour research.

In the mid-term, after repeated exposure to AI-manipulated interactions, observers may learn and recalibrate the likelihood distribution $p_{online}(\mathcal{O}/\mathcal{H})$ based on regularities observed in the filter-pervaded online environment. All bets are off, in truth, about how these distributions may shift. In Western cultures where smiling is highly desirable, online smiles may end up losing all predictive value if filters are used “all over the board”, regardless of the person’s actual mental state; on the contrary, smiles and other expressions may acquire an inverted value if we learn that such filters are used predominantly to hide unpleasant affect: with a cheering prototypical AI-smile or voice, our Paula may in fact be indicating that she’s perhaps not feeling very well. Should facial and vocal cues so lose or change their long-developed informative value, consequences on our social cognitive abilities will likely be profound. It appears especially important to research how uncertain, AI-manipulated expressive cues may affect the development of social cognition in younger online users, as the critical window for developing such abilities largely overlaps the age-group with the most intense online-platform usage [7]. This situation is of unprecedented importance, because emotions- constructed from these clues and cultural environment- are collective cognitive tools essential to the proper running of our society [8].

Spread of the ”Tiktok likelihood function”

In the longer-term, releasing the biological constraints that currently hold on mental-state expressions in the face and voice may also be associated with hard-to-predict effects on the cultural dynamics of social behaviour.

First, it is poorly understood whether priors and likelihood distributions learned online and adjusted for the possibility of technological control -let’s call it the “Tiktok likelihood function”- would remain circumscribed to the digital world, or instead generalise to the offline world. In the first instance, we may see the emergence of alternative cultures of emotional expression (as e.g. with emojis) , that do not merely mimic those found in real-life interactions; but co-exist with them. In the second, in our view likely, instance, we may see the emergence of what one may call “online-centric errors” where a smiling Paula met on the street will be thought sadder than she really is, based on the smile she might have showed if we had met her online (i.e. on

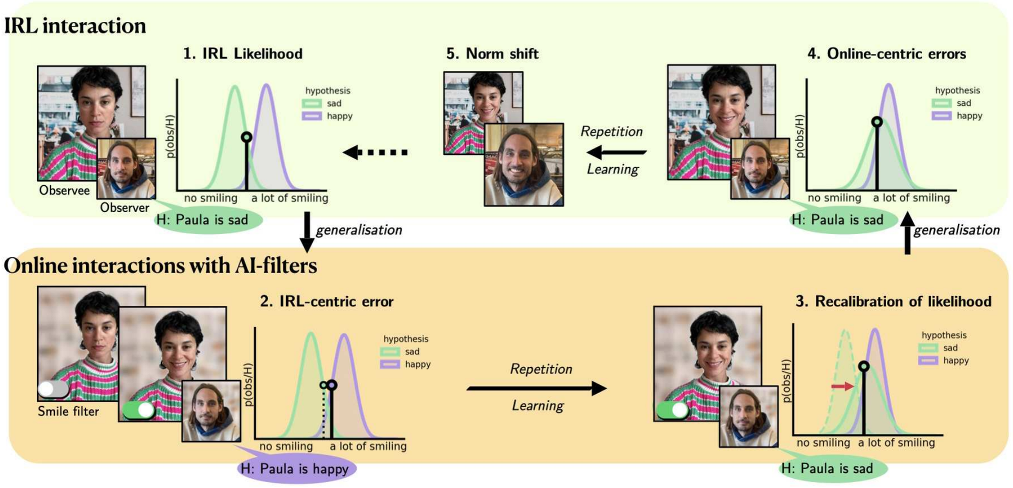

Figure 1: **How AI face and voice filters could reprogram social inferences:** because we infer each other's mental states based on regularities learned from previous interactions, the adoption of AI-filters of social expressions will alter the parameters of our everyday social cognitions both online and "in real-life" (IRL). **(1)** Meet our friends Paula and Jules in a Parisian café. According to Bayesian inference, Jules may infer that Paula is a little sad today based on his previously-learned likelihood distribution $p(\mathcal{O}|\mathcal{H})$ to observe Paula's particular facial expression when she is sad or happy. **(2)** If they meet online instead of IRL, Paula may be using a filter to control her facial expression independently of how she really feels. By modifying the observation $\mathcal{O}$ within an otherwise unchanged probability distribution, this may lead Jules to conduct an "IRL-centric error", falsely concluding that Paula is in fact happy. **(3)** After repeated exposure to AI-manipulated interactions, Jules may learn and recalibrate his likelihood distributions, such that e.g. a smiling face now indicates that Paula is in fact hiding a low mood. **(4)** Likelihood functions recalibrated online may then generalise to unfiltered IRL interactions, leading to "online-centric" errors where Jules may take a smile he'd have previously interpreted as genuine to instead mean sadness. **(5)** Because perceptual expectations also shape one's own behaviour, this process may in turn lead to shifting social norms for how facial expressions are valued and used - for instance, if it becomes common to use a filter to hide one's low mood online, Jules may feel pressured to smile more IRL too. In sum, because they algorithmically release the biological constraints that have so far held on the cultural evolution of facial and vocal expressions, AI-filters have potential to affect our social-cognitive processes and, over time, human sociability in general.

the computation of $p_{online}(\mathcal{O}_{offline}/\mathcal{H})$). In this regard, the use of beauty filters seems to already impact the behaviour and mental state of users beyond the digital ecosystem itself. Several phenomena have been documented such as selfie dysmorphia, defined by self-esteem problems and body distortion among regular selfies takers using filters [4], and issues leading to an increased recourse to cosmetic surgery in order to enhance its online appearance. Factors favouring the spreading of "Tiktok likelihood functions" to our offline social inferences may include both personal factors such as emotional traits and amount of filter usage, but also technological ones like the development of more immersive and realistic online environments. Studies should explore the benefits of pedagogic programs designed to educate users on the cognitive effects of these social-affective artefacts.

Second, while moral values are known to influence the use of technologies [9], technology use itself also has the potential to exert changes on our moral landscape, described as *soft impacts* of technologies [10]. Because of filters, non-verbal expressions which were once accepted as inevitable (e.g. a slight tremor in the voice when one's nervous) may become controllable, and thus blameable and subjected to explicit or implicit social coercion (e.g. *"why didn't you put stress-control on?"*). This "norm shift" outlined for augmentation technologies, seems equally applicable to these AI-filters.

Conclusion

In sum, because human minds infer each other's mental states based on regularities observed in their environment, the adoption of AI-filters of affective expressions in online interactions could alter not only our outward appearance but also, and above all, the internal models on which we base our every day social inferences and, over time, human sociality in general. Here, we propose Bayesian predictive inference as a framework for examining this anthropotechnical potential —i.e., the shaping of our social interactions— of these devices. Each of these hypothesis can be investigated with empirical research through observational data (e.g. records of filter use on social media and behaviors) or cognitive science experimental design (e.g. giving interacting participants the possibility to use filters, or not and assess their inferences). Crucially, studying how AI filters might reprogram our social cognition and sociality could yield invaluable insights not only into the theoretical models that helps to explain our physical and digital interactions - and how they might differ, but also to inform an "ethics by design" methodology for AI affective artefacts. Finally, such reflection would contribute to the necessary interdisciplinary critical thinking commensurate with the philosophical, scientific, and societal challenges that these tools pose to our digital society.

References

[1] N. Ambady, R. Rosenthal, Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis., Psychological bulletin 111 (2) (1992) 256.

[2] K. J. Friston, C. D. Frith, Active inference, communication and hermeneutics, Cortex 68 (2015) 129–143. doi:https://doi.org/10.1016/j.cortex.2015.03.025.

[3] L. F. Barrett, Emotions are real., Emotion 12 (3) (2012) 413–429. doi:10.1037/a0027555.

[4] A. Javornik, B. Marder, J. B. Barhorst, G. McLean, Y. Rogers, P. Marshall, L. Warlop, 'what lies behind the filter?'uncovering the motivations for using augmented reality (ar) face filters on social media and their effect on well-being, Computers in Human Behavior 128 (2022) 107126.

[5] A. Clark, D. Chalmers, The extended mind, Analysis 58 (1) (1998) 7–19.

[6] R. Tucciarelli, N. Vehar, S. Chandaria, M. Tsakiris, On the realness of people who do not exist: The social processing of artificial faces, Iscience 25 (12) (2022).

[7] A. Ferguson, G. Turner, A. Orben, Social uncertainty in the digital world, PsyArXiv m97ug (2023).

[8] J. Searle, Making the social world: The structure of human civilization, 2010.

[9] N. Guerouaou, G. Vaiva, J.-J. Aucouturier, The shallow of your smile: the ethics of expressive vocal deep-fakes, Philosophical Transactions of the Royal Society B: Biological Sciences 377 (1841) (2021) 20210083. doi:10.1098/rstb.2021.0083.

[10] S. van der Burg, Taking the "soft impacts" of technology into account: broadening the discourse in research practice, Social Epistemology 23 (3-4) (2009) 301–316.