AI and Ethics
https://doi.org/10.1007/s43681-023-00372-8

ORIGINAL RESEARCH

# Socialisation approach to AI value acquisition: enabling flexible ethical navigation with built-in receptiveness to social influence

**Joel Janhonen**[1,2]

Received: 24 August 2023 / Accepted: 24 October 2023


**Abstract**
This article describes an alternative starting point for embedding human values into artificial intelligence (AI) systems. As applications of AI become more versatile and entwined with society, an ever-wider spectrum of considerations must be incorporated into their decision-making. However, formulating less-tangible human values into mathematical algorithms appears incredibly challenging. This difficulty is understandable from a viewpoint that perceives human moral decisions to primarily stem from intuition and emotional dispositions, rather than logic or reason. Our innate normative judgements promote prosocial behaviours which enable collaboration within a shared environment. Individuals internalise the values and norms of their social context through socialisation. The complexity of the social environment makes it impractical to consistently apply logic to pick the best available action. This has compelled natural agents to develop mental shortcuts and rely on the collective moral wisdom of the social group. This work argues that the acquisition of human values cannot happen just through rational thinking, and hence, alternative approaches should be explored. Designing receptiveness to social signalling can provide context-flexible normative guidance in vastly different life tasks. This approach would approximate the human trajectory for value learning, which requires social ability. Artificial agents that imitate socialisation would prioritise conformity by minimising detected or expected disapproval while associating relative importance with acquired concepts. Sensitivity to direct social feedback would especially be useful for AI that possesses some embodied physical or virtual form. Work explores the necessary faculties for social norm enforcement and the ethical challenges of navigating based on the approval of others.

**Keywords** Machine ethics · AI value acquisition · Socialisation · Conformity · Social intuitionism · Embodied learning

## 1 Introduction

Artificial intelligence (AI) uses computational problem-solving to process information and has succeeded in performing an expanding variety of tasks. These have come to include numerous real-world and social applications in addition to being used for statistical modelling. This article explores questions of Machine Ethics—a field concerned with the ethical behaviour of artificial agents [1]. Machine ethicists generally focus on abstracting principles from observed human behaviour that could assist in guiding AI's operation. Exact rules may be great in theory but have limited utility for the complexities of life. High-minded ethical ideals and codes appear closely useless in technological practice, which has been argued to qualify AI ethics as a misuse of attention and resources [2]. As explored ahead, such an approach significantly differs from how humans make decisions and navigate their way in the world. Factors that best explain human behaviour relate to environmental interplay and practical social organisation rather than abstract principles or context-independent rules. Generalisations only take an agent so far without the ability to directly pick up situational cues. Some researchers engaged in Human–AI interaction have advocated for imprinting AI's "genome" with the necessary social intelligence to manage complex human situations [3]. The present work takes the appeal to socialise AI seriously by exploring the necessary social ability and its connection to value learning and overall moral development. Regarding the design of intelligent and autonomous machines, this

✉ Joel Janhonen
joel.janhonen@bioetiikka.fi

[1] University of Turku, Turku, Finland

[2] Finnish Institute of Bioethics, Tampere, Finland

Published online: 21 November 2023

work endorses the paradigm of embodied AI [4] but with a special emphasis on emotion recognition and social learning.

Compatible functioning is crucial within shared environments that are not free of risks. To avoid unnecessary harm and trouble through joint effort, parties need to be constantly aware of communicated risk signals, the feelings and intentions of others, and points of tension. Agents who cooperate or merely coexist in the same environment can greatly benefit from being receptive to information relayed through a variety of means. Much of the knowledge that contributes to our decision-making is generated by affective reactions, which are an outcome of psychological adaptation to the demands of the environment. These can be thought of as visceral flashes of negative or positive feeling [5]. Even our moral decisions are greatly influenced by what we feel arising within us and what we sense arising in others. These affective judgements can be instantly formed even about the most unintelligible things. Real-time collective action requires intuitive navigation ability, providing the group with a shared sense of direction. As opposed to communicating everything through language alone, emotional signalling functions to reflexively transmit these affective insights.

Many of these non-verbal expressions are human universals and others are a product of sociocultural processing [6]. Verbal communication that has a more logical structure is a rather late addition to information exchange, evolutionarily speaking. Even among humans, not all that is relevant can be conveyed in unambiguous statements. When dealing with "straight facts" is insufficient, signals of displeasure or disapproval can convey useful information, even without rational exchange or the ability to articulate the concern. This work uses the term social signalling to cover all human modalities of non-verbal communication but with an extra emphasis on messaging that expresses a volume of attitudinal favourability. This is considered a social reference, a useful indication of a judgement. Here, a message simply means information to be transmitted; signal instead is the form it takes during its transmission.

This work explores the creation of socially sensitive AI—both in terms of improving user experience and contributing to the making of moral decisions. The aim is to demonstrate that there is much more to morality than rationality and so highlight the importance of exerting normative social influence on any agent operating within a shared environment. This is done by exploring literature on the proposed basis and function of normative judgements and conformity. The described alternative approach to AI value acquisition intends to evade the need to specify a finite list of universal norms, principles, or values [7]. Instead, the appropriateness of actions is estimated directly from user interaction. Receptiveness to social references is entertained as a solution for AI alignment. Rather than attempting to embed a comprehensive ethical code into AI's algorithm, sensitivity to social norm enforcement could function as a domain-flexible capability to broadly guide artificial judgement formation. Making individual and group reactions intelligible for AI opens new data streams that include knowledge about the shared norms and associations of importance. Yet, moral navigation based solely on the avoidance of human disapproval gives rise to important challenges that are discussed in detail. Apart from the general usefulness of conformity, it can also perpetuate less-adaptive or unwarranted conduct.

Directly or essentially mimicking the mechanisms that are responsible for normative conformity among people could enable a kind of AI socialisation. The necessary undertaking that enables these systems to learn how to behave appropriately within society may not fundamentally differ from the general trajectory of human moral development. The human trajectory is bottom—up—value learning rather than loading—and heavily reliant on key social capabilities. New concepts may be accreted only on top of the existing ones and gaining a sense of their importance takes time and trial. Values that have emerged from advantageous patterns of human behaviour may only be internalised through embodied learning and active social involvement. This work mostly discusses norms and morality in pragmatic terms, meaning that ethical notions are examined for their context-dependent usefulness, rather than perceiving them as universal truths. This viewpoint is well suited for a discussion on AI value acquisition, given that the objective is to equip it for appropriate and collaborative conduct, not to perfect moral philosophy.

### 1.1 Artificial agents

The primary focus of this work is on AI systems that can be thought of as artificial agents. This can be understood as things designed for an active role, even extending to the making of morally significant decisions. Successful functioning in the real world—i.e., within an open environment not designed for a specific application—is not attainable with pure logical problem-solving, but by resolving the system–environment interaction [8]. When attempting to understand and replicate aspects of human cognition, it is a mistake to reduce it to logic or computation and so dismiss intuition and embodied intelligence that hinges on the body plan and physical constitution of the agent. Reasoning and intuition—including emotional processing—are here both perceived as cognitive functions. Useful knowledge can be generated regardless of the type of information processing, responding to different environmental demands. The unconscious pattern synthesis of intuitive cognition has much to do with enabling rapid decision-making that is grounded, situated, and hence suited for real-world choices [9]. More analytical thinking grants additional capabilities to process information, but it is not the core of our cognition.

The active role of an artificial agent does not need to be anything miraculous or comparable to real persons. For the socialisation approach to have relevance, an AI application only needs to operate within a social context with a non-negligible degree of independence. In the case of embodied applications, where a computer program controls a physical or virtual form, these roles can involve performing varying logistic, security, and service tasks that require some social exposure. Instead of mere models, these would be systems with real-life implementations and direct outcomes. Possible embodiments of AI include anything from a vocal interface to an on-screen character or a simple robotic limb. All that is required is some tangible characteristic or expression that makes it part of the observable world. A unique embodiment creates a unique role or system–environment interaction.

Agents do not only passively receive input from the environment; rather, they actively direct their "gaze" and enact in relationship with it. In a social environment, the primary interactees are other agents. Almost any tangible form that can be perceived could be used for social signalling and hence produce a response. Even amazingly simple movements of two-dimensional geometric shapes have been demonstrated to successfully convey emotions and intentions [10]. An agent that can participate in two-way communication can be argued to have social ability in addition to mere reactivity [11]. Any claim over the communicative space requires sensitivity to transmitted signals and awareness of the social forces present. Conversely, AI models developed to allocate resources, perform data crunching, assist in the design, or otherwise only deal with "straight facts" would not benefit from being susceptible to direct social influence. Here, employing and excelling in logic are sufficient. This article concentrates on the receiving side of social interaction, yet it is good to remember that for many real-world applications, communication would flow in both directions.

### 1.2 Intangible human factors

AI systems that have more general abilities and operate in society among people should have an algorithmic architecture that accounts for a requisite set of factors. As the abilities of these systems get broader, their programming should correspondingly account for a more complete set of ethical considerations, many of which have complex social foundations. One does not need to view AI as an existential threat to prefer an artificial interactee that approximates appropriate behaviour over one that neglects either explicit rules or basic assumptions. Once an agent's sphere of influence spreads to new areas of interaction, there are an increasing number of factors that need consideration to function properly. In the extreme, a hypothetical general AI should consider all our interests, resolve conflicts between them, and align its operation accordingly. Mastering any new skill or learning any concept requires prerequisite knowledge and ability, much of which is too automatic and commonplace for us to even consider. As social creatures, a significant portion of this basic ability is social.

When operating in the human context, these considerations do not only cover tangible factors of the physical world or easily quantifiable conceptual units like money or seconds. Instead, they encompass varieties of social factors that defy defining, let alone exact mathematical modelling. The formulation and meaning of these concepts as well as the degree of associated importance also vary between cultures and individuals. Even for agents like us, who have specifically evolved to apprehend this space, the ambiguity of social signals, standards, and tensions can at times leave us clueless and embarrassed. Yet, we are generally able to account for these intangible factors in our decision-making even without completely understanding them in rational terms. There is a significant degree of automated processing taking place, largely on the emotional level. Positive and negative flashes of affect, either arising within us or transmitted to us, push and pull us toward more appropriate directions. Especially regarding social decisions, our emotions help by guiding interpersonal decision-making [12]. Beneficial emotional drivers of decision-making automate how we account for many intangible factors without increasing the cognitive load.

### 1.3 Counter-intuitive solutions

Digital ethicists and philosophers have called attention to a prospective threat regarding AI systems with more general problem-solving capabilities. The concern is that such systems could pursue self-conceived instrumental goals with undesirable and unforeseen consequences for humanity and its priorities [13]. A super-human ability to reach objectives without a complete appreciation of our priorities could undermine human well-being if the steps taken between the initial state and the set goal would inadvertently suppress something precious to us. No matter how great, good, or harmless the objective is, the causal pathway should not be paved with unintended consequences. To find an optimal solution for a given task, an artificial agent would need to cross-evaluate the extended outcomes of all hypothetical intermediate actions. In the pursuit of its final objective, any direct or indirect impact on humans—or other living things—should be accounted for and priced in the decision-making. The clear impossibility of performing a comprehensive analysis on every prospective action in relation to everything that matters to us well illustrates the need to rely on approximations to guide operation. Not unlike how humans manage to navigate and collaborate amid all this complexity without experiencing non-stop cognitive overload.

To prevent actions or inactions that seriously conflict with the goals and priorities of humanity, there is a general understanding that the goal system of AI should in some way represent our ethical operating system. Our decisions are guided by our values; hence, these algorithms should be made to share these underlying priorities. A value here simply means an association of importance, either with or without an unambiguous or programmable description. Misalignment of underlying values would result in misaligned navigation and a threat of drifting to a collision course. Moreover, if any relevant consideration is absent in the programming, it will simply be ignored, regardless of how self-evidently important it would be to us. Without specifically designing AI with the capacity for intuitive or common-sense responses, these will be missing. The chance that there exists some unrecognised, unappreciated, or mathematically indescribable yet crucial factor that needs to be accounted for is not zero. Decisions made without their weight risk being counter-intuitive to our moral sense. Moral psychology, which studies human decision-making and moral development, should be consulted whenever considering the internalisation of corresponding values into artificial minds.

### 1.4 Aligning intention

The developers of AI have encountered great difficulty when attempting to program human values into algorithms, which has been termed the value-loading or acquisition problem [13]. Top-down approaches that seek to ensure ethical behaviour by trimming down the decision tree with rule sets have by themselves proven largely insufficient when encountered with the complexity of our world [14]. Dictating an explicit list of values that would come close to a complete representation seems impractical or unthinkable. Therefore, designing dynamic methods for conserving and correcting the alignment of priorities and behaviour is highly desirable, preferably while AI's capabilities remain limited. Otherwise, it is easy to see how misalignments could emerge between logic-based systems and our complex and often implicit systems of subjective and intersubjective value. For now, the definition of alignment is imperfect, and we lack the means to measure it. However, the clarifications provided by the Alignment Research Centre help to illustrate this challenge. Accordingly, to achieve alignment means that the goal-directed behaviour of AI aims to meet human preferences, or what we think is the right thing [15]. Therefore, intention-aligned AI can still make mistakes, but it "tries" to act according to our wants and meet our standards. This fits well with the idea of utilising social conformity and norm enforcement in AI value learning.

The sought-after AI value acquisition is additionally complicated by meta-ethics; namely, arguments and observations that the moral interface of humanity is not founded on abstract principles or absolute values but on practical demands, moral emotions, and relational influencing. Reasons to justify decisions are generally formulated only after the judgement itself has been formed based on intuition and prevailing emotional dispositions [16]. Even if assuming that ethically salient factors could be accounted for via a rule-based approach, managing human situations with idiosyncratic personal concerns or group dynamics requires a social skillset rather than a directive. Theory-vice, this article leans on social intuitionism to highlight the non-rational character of moral decision-making and the precondition of social ability—concerning any agent learning to act in a conventionally appropriate way. Morality is perceived as the art of living together. Social concern and moral dispositions are understood to enable this reliable collaboration and so possess adaptive evolutionary value. The additional need for moral reasoning and the inclusion of categorical imperatives into algorithms are also discussed. Overall, the work argues that the capacity necessary for being a socially intuiting agent needs to be built in, including a goal system with prosocial characteristics. If this capacity is sufficient for the agent's context, embodied form, and life tasks, value learning can follow via socialisation.

### 1.5 Modulating behaviour

The problem of value alignment follows from the fact that AI does not inherently share the problems or priorities innate to us. Its core motivations—like decreasing the value of the error function or increasing the value of the utility function—are worlds apart from the essential priorities of higher animals. Even though neural networks are made to resemble the circuitry of a human brain, artificial minds radically differ in how and why they process information [17]. Judgements we consider to be obvious and might not even bother to articulate like the undesirability of enduring pain or boredom are not self-evident for these agents.

Despite this dissimilarity, AI can nevertheless be directed to pursue more familiar objectives by tuning its goal system, meaning that a mathematical reward is received from completing a specified task. One such goal could be to avoid actions humans would disagree with, which would require it to be educated on the desirability of different actions. As a simple application, a social error function—i.e., the difference between the desired and the performed action—could be derived from user feedback. With this sensitivity, AI could be trained on data that adequately reflects human judgements in specific situations. In theory, this could instruct it to operate "desirably". Something like a social error function that enacts a sway over the goal system could help align the intentions of an artificial agent with its interactees. However, could an algorithm be designed to sufficiently mimic caring about our values to appropriately

function in a shared environment? Even if deep down its preferences over outcomes would be based on nothing more than an arithmetic function. Additionally, could training data gathered from multimodal user feedback contain enough information about the norms attached to versatile and dynamic social roles?

Systems that make morally relevant decisions need a goal system that effectively approximates our own while being quite unlike it. From the hierarchy of human needs, motivations to seek affiliation and belongingness relate to our high natural sensitivity to cues of social rejection [18]. Even if our physiological needs would be quite distinct from a robot, the inherent need for social approval should be shared by both. Our basic needs and resulting motivations do not directly dictate our actions. Instead, interactions with and within the context greatly inform behaviour and decisions. Dependency on social inputs makes the human goal system externally malleable. The same human hardware can be differently programmed—namely through socialisation. Yet, this process of learning and internalisation depends on requisite biological and social capabilities.

Many action-modulating layers can form between the primal drives and resulting behaviour, given that the opinions and feelings of others are intrinsically regarded as important. The history of humanity demonstrates that natural motivations are flexible for modulation and can give rise to vast varieties of behaviour, value, and social organisation. Given that our inherent animal motivations can be socially directed to pursuits of higher value and so enable us to adapt and integrate to vastly different normative environments, the toolkit of socialisation may do the same for AI systems by shaping how their motivational system (i.e., components of the utility function) is mapped to specific concepts and actions. At least from the viewpoint of substrate independence, simple artificial drives to attain mathematical rewards could sprout comparable complexity given effective social modulation [13]. For an artificial agent, the hierarchy of needs could likewise consist of similar motivational layers of neural circuitry, built upon one another [18].

### 1.6 Shared notions of importance

Our physical constitution and life-plan shape and set limit to our conceptions of the desirable. The desirability of concrete objects or actions is largely shaped by natural considerations like eating, sleeping, and mating. Agents that have nonadjacent basic needs and motivations can nevertheless sustain compatible behaviour by having the necessary social ability and sensitivity. The resulting interaction can hence bring about joint efforts toward aligned goals. Collaboration or prosocial behaviour among agents is not a necessary outcome of shared notions of value; even identical core motivations can evidently produce misaligned instrumental goals and conflict among people. Practical alignment of life tasks and a spirit of collaboration may even be more important than matching values; getting along is more about social ability. Instead, the shared physiological motivations of animals continuously play out their never-ending conflict in nature, generally even without a possibility of arriving at a collaborative social understanding. Meeting the threshold for compatible and beneficial coexistence requires the skills and tendencies suited for it. Therefore, regarding any prospective collaborator, it is desirable that their motivational system is susceptible to the social sway of our species.

### 1.7 Embodied social learning

In the field of pedagogy, the *Embodied learning* approach gives priority to affective and sensorimotor experiences rather than focusing mainly on mental factors [19]. "Bodied beings" who are situated in an environment attain practical knowledge through repeated interactions with the elements of the environment and other "bodied beings" within it. Knowledge construction is thus seen both as a physical and mental act that needs the body as well as the mind. This view, that learning is achieved through interaction with the environment and by sketching and interrogating representations of it, is also gaining popularity among AI researchers [20]. Yet when exploring how humans make sense of objects and changes in their environment, neither affective nor ontological information appears to have the overall processing priority, but their order instead depends on the context [21]. It would be useful to know which factors make us prioritise affective processing in certain social situations and suppress it in others. By adjusting the priority given to emotion-evoking social signals, AI's responses to stimuli could be made more flexible and context-sensitive. This way, an embodied artificial agent would get the most out of the key experiences during its embodied value learning process. To emphasise, these learning events are primarily about diverse social exposure [22].

Researchers behind an article that aptly discussed the "dark matter" of AI pointed out that the common perspective of intelligence lacks an important insight about human cognitive architecture [23]. They propose that interpersonal intelligence—cooperating well as a whole—is tightly connected to intrapersonal intelligence. The former is considered to enhance the latter through social learning mechanisms; knowledge is best acquired and applied by many agents working together. At least for humans, the adaptive modulating biases that are adopted based on social interaction are necessary elements for our advanced cognitive ability. Apparently, this is still a rather unfamiliar view in the field of AI and so social interaction merits the description of "dark matter" regarding the picture of intelligence (also see

social learning strategies and recent work on communicative agents and multiagent systems).

Natural examples of collective intelligence hint at the possibility of generating flexible means to optimise behaviour, where interaction between many individuals gives rise to emergent phenomena. Besides being necessary for appropriate behaviour, social interactions may even be essential for advanced cognitive ability. For more on research regarding embodied artificial social agents, see the full perspective by Bolotta and Dumas [23]. The present work also entertains that mechanisms of human social learning can inform the development of AI by, e.g., allowing efficient transmission of information from human-to-AI. However, the creation of artificial social ability is discussed primarily as a method for value acquisition.

### 1.8 AI socialisation

To successfully engage in dynamic social scenarios, artificial agents would require some degree of social intelligence [24]. This, in turn, presupposes receptiveness to detect and the ability to process social messages. This work explores how AI systems could be trained on non-verbal cues as instructions to induce a change in their conduct. AI systems should learn to prioritise or limit themselves to actions that are considered acceptable, right down to each sub-goal of the procedure. As social AI learns to reliably discern expressions of sadness, joy, disgust, and anger from different streams of data, these factors should influence its decision-making. Such sensitivity to social forces is distinct from social intelligence and should be given priority as explained ahead. An appropriate response to forms of communicated displeasure and disapproval would enhance the user experience, safeguard against inappropriate solutions, and possibly even build rapport.

Since an all-encompassing set of rules or a universal ethic does not exist, the ability to "read situations" provides a basis for appropriate social navigation. AI that is designed to be receptive to emotionally communicative reactions could learn to deduce whether they convey acceptance or repudiation. Social references about the situational desirability of actions can be then used to learn what is appropriate and when. Our social signalling is filled with normative subtext. By mapping which reaction generally stems from which word, behaviour, or situation, AI may organically learn to pre-emptively steer away from trouble by aiming to minimise expected disapproval.

As described before, the value-loading problem is the challenge to sufficiently embed—often complex and contradictory—human values into AI. The socialisation approach to the value-loading problem is to make the algorithm submissive for social norm enforcement by embedding sensitivity for approval and disapproval into its goal system. The approach can be divided into two challenges: first, the interpretation of cues and, second, the dependency on the approximated level of approval when deciding on an action. A conformist AI would be designed to intuitively consider the sensitivities of its interactees. These inferred concerns could then be prioritised over self-set or even externally given objectives. This would help to ensure that the AI system would act within the limits of what is considered appropriate, given that it "reads the room" well enough. The reading of social references may also enable the internalisation of values expressed by them.

This work proposes that the creation of moral faculties in AI systems could be delineated in terms of socialisation and conformity. AI-powered agents that read social signals, discern context-specific precursors for actions and reactions, as well as model them in relation to cultural factors, could be trained to make decisions that reliably conform to societal norms. Throughout our daily interactions, billions of potential data snippets are disseminated across the communicative space—encompassing the echoes of our expressed moral positions. Despite some of the data reflecting ill-intent or other human moral shortcomings, these judgements portray the considerations involved in the conventional decision-making. Successful socialisation might not depend on experiencing moral emotions or having a human-like motivational structure. It may be sufficient that an agent is receptive to expressed normative attitudes and has processes to generate useful knowledge from them. This requires the motivational system to be open to the transmitted affect. Designing sensitivity to affective insights could equip AI with a substitute for our moral emotions, enabling it to navigate social life and embrace the moral code of its context. The following exploration of the connection between human sociality and morality aims to provide a theoretical background for the AI socialisation thesis.

## 2 Social prerequisites for morality

Socialisation is important for human moral development. Through this process, individuals learn what is considered appropriate and desirable in view of the role they are destined to fill in society [25]. When successful, it makes interaction and collaboration possible in the shared social environment. Value learning is a key part of this process, resulting in the internalisation of norms, motives, and beliefs of one's family, peers, and the larger society. Simply put, after picking up a new concept, a level of relative importance becomes associated with it. Rather than reason or logic, this primarily happens through interpersonal identification and interactions saturated with explicit and implicit signals of approval [26]. Socialisation allows individuals to learn the values specific to their culture and the ways of behaving

that receive approval and acceptance rather than condemnation and social punishment. These considerations are not only acquired as knowledge, but they become a part of one's nature. Therefore, socialisation is essential for developing one's sense of right and wrong, making these judgements automatic. The intuitive nature of such judgements is supported by a finding that people manage moral tasks even when carrying a high cognitive load. Automating moral responses with intuitive-emotional processing decreases the need for continuous and intense cognitive processing or principled reasoning in social scenarios [27].

Until now, the process of socialisation has been the only way to make an agent that acts in an approved manner. A hypothetical alternative method of directly "downloading" or encoding norms into an agent would be a radical divergence. Because it is not obvious that all normative content is codable, a process analogous to socialisation could also have relevance for silicon-based decision-makers. Such a learning process could enable the creation of functional responses to human social signalling that are within culturally sanctioned boundaries. However, this only becomes possible if the AI's algorithm and goal system are made sensitive to social forms of influencing. Concern over negative feedback needs to be built in, as well as the means to detect it. The current sociological theories should inform AI research about the process of social influence, so that it can be used to transmit complex cultural elements that simply cannot be loaded in.

Types of non-verbal human communication differ in meaningful ways. It can be verbal, voluntary, intended, conscious, or none of the above. The list of physical cues includes the positioning of the eyes, facial expressions, body language, and the use of distance. This is a major part of human social interaction and embodied AI systems need to be able to read and respond to it to become good interactees. Transmitted signals that provide information about social facts may indeed divulge much about our external and internal states [6]. Yet, when discussing value acquisition, the focus should be on the communicated normative attitudes or disapproval rather than each affective judgement. Negative flashes of feeling that are caused by an everyday disappointment should be differentiated from ones indicating mistreatment or actual social misconduct. Dispositions to feel a certain way about specific actions are explored ahead in connection to human prosociality and moral development. But first, a short section on conformity.

### 2.1 Normative conformity

The signalling of disapproval and approval through communicative social channels has existed within humanity far longer than codified norms or abstracted moral values. Before a cognitive moral capacity can form, the initial normative stances need to be "soaked up" from the social environment. This work understands conformity as a condition of being aligned with established conventions and hence as an outcome of successful socialisation. Not just obedience to the given limitations but proper internalisation. In psychology, this term is used to describe the general tendency of people to match specific behaviours, as demonstrated by the famous Asch conformity experiment [28]. Instead, this work mainly considers normative conformity. Rather than what can be called "social lubrication"—like spontaneously laughing with others at a joke you did not understand—normative conformity is the pressure to fall in line with socially held value judgements. This tendency to match judgements and conform to norms is here primarily considered a moral positive and a prerequisite for any complex social organisation—although with important caveats that are explored later. Contrary to conformity, enforced compliance can also motivate collaboration. Compliance does not require norm internalisation, so mutual agreement is not necessary [29]. In a compliant social group, the enforcement mechanisms wielded by established authorities pressure individuals to stay 'in line'. However, to attain reliable collaboration, there also needs to be internal pressure that promotes it.

### 2.2 Evolved to be socially conscious

Effective means of behaviour modulation enable more elaborate social organisation, which becomes apparent when exploring the history of our species. In his article on the emergence of moral demands, Matteo Mameli lays out a hypothesis about the role selection pressures have played in shaping our valuative attitudes. Accordingly, practical demands for reliable collaboration sometime in the Late Pleistocene created emotional dispositions against non-collaborative behaviours [30]. Thus, perceived violations against these demands came to spark negative feeling, whether witnessed or personally considered. The evolved tendency to feel this spectrum of moral feelings is here thought of as conscience, which functions to circumvent the anger, disgust, and disapproval of the social group. Our normative attitudes would therefore be an evolutionary adaptation that enabled aligned behaviour and broad collaboration. Innate emotional dispositions transcend status hierarchies by creating authority-independent norms. As a result, members of the group conform to social standards by following them internally, as opposed to being compelled to comply by more powerful individuals. Notably, our evolved dispositions can prioritise the in-group at the expense of outsiders, which is why our innate prosociality requires to be extended. The important difference between superficially compliant behaviour and norm internalisation will be revisited in the later discussion.

Socially (i.e., non-genetically) passing-on patterns of behaviour that have been found generally successful—or

"good"—have certainly increased humanity's adaptive and collaborative capacity significantly. When the natural ability to intuit judgements is coupled with the cultural transmission of values and useful patterns, the result seems to be a collection of efficient problem-solving methods [31]. Utilisation of this heuristic toolbox reliably results in adequate solutions; like how moral intuition can give insights to address obstacles of collaboration. Prosocial behaviour proved evolutionarily advantageous even without an expectation for immediate or equal reciprocity. Oppressors and cheaters who are thought to hamper cooperation and create conflict reliably evoke and receive disgust and cues of condemnation. Even in the absence of material or immaterial sanctions for wrongdoing, the perceived legitimacy of these normative expectations can still motivate a significant number of people to comply with shared norms [32]. This motivation can hold even if a personal cost is adjacent to adherence. We are emotionally predisposed to seek to fulfil normative expectations. This intention needs to be passed on to any artificial social agent by designing an artificial conscience that manages to circumvent grounds for social punishment.

### 2.3 Reprogramming moral intuition

It appears that what caused us to originally advance in the art of living together was an elaboration of our motivational architecture. According to the modern synthesis of moral psychology, our basic evolutionary motivations—to either "avoid or approach"—are connected to the base of our morality [5]. Especially regarding a gut feeling like revulsion, it is easy to see how our affective and normative judgements are linked. Emotions like anger and disgust came to play a moral role as they were directed toward perceived norm violations through adaptive social modulation. Rules are preceded and backed by moral motivations. Our sense of right and wrong is largely nested in socially binding relationships, all with their specific expectations. These connections and their accompanying forms of social persuasion function to keep our less-desirable qualities in check and embed us into a fit social fabric.

As we evolved to require a tight and dependable social group, the patterns and reactions that promoted collaboration were passed on through social conditioning. For the present work, this is understood as the mechanism that provides an individual with the necessary means to form socially appropriate judgements. Classical conditioning is a form of enforcement learning, i.e., a process that directs the mind to evoke a specific response to a specific stimulus. An affective reaction can also be triggered via social means, like through a transmitted threat of punishment. Hence, this psychological capacity can be converted into a sociological training process. Repeatedly triggered responses can over time tune an individual's brain, so that instead of performing an antisocial action—just entertaining it is sufficient to unleash the negative feeling [16]. To enable intuitive social navigation, such affective markers are integrated with the agent's knowledge and planning functions. Within our brains, emotional processing and decision-making are banded together in the ventromedial prefrontal cortex. Channelling of negative emotions to specific actions may be the essence of socialisation, from emotional dispositions to held positions. As a result, our moral intuition conforms to what is deemed appropriate in each context.

Evolving practical demands of the environment require periodic readjusting of behaviour. Consequently, emotional dispositions and norms also need requisite flexibility to be revised and remain adaptive. In the modern context, anger and disgust have been directed toward present-day violations like treason, racism, or tax fraud. Yet, before adequate ability in conceptual thinking, attaching negative reactions to more abstract violations is impossible. It is difficult to express anger toward an insider trader without understanding how a plethora of relevant concepts relate to each other. Before grasping these, one needs to go by social references. Neither compliance nor conformity is absolute in humans, but these are balanced with the propensity to oppose and challenge which at times manages to redirect our collective feelings about an issue. Even if we have evolved to generally be swayed by social forces, we also have the capacity to resist them. Agreeableness and compliance are not automatically better strategies than defiance. However, the likelihood that an individual has a better stance than the broader social group and so should resist the pressure to conform only increases with a better comprehension of the world. Therefore, critically assessing and detaching from the shared perspective is arguably situated later in the trajectory of moral development [33]. Critical assessment of social standards relates to the post-conventional stage of moral development—discussed later. Moral awareness is the capacity to think normatively, not merely to act based on unrecognised dispositions but to detect and appreciate competing ethical aspects. Yet importantly, at least for humans, the ability to critically reason follows the ability to conform.

Being inherently motivated to avoid social discomfort and negative feedback is important for becoming a socially able agent. Biological discouragement of certain undesirable actions can be exemplified by blushing out of shame, a unique capacity of the human species [34]. The involuntary transmission of this social signal is a great example of how we are selected for prosociality. The shame response and other alike biological features of our moral sense demonstrate how the socially enforced programming is imprinted into our biological hardware. Things that trigger shame are also dependent on social context. Even if a disposition to feel or react is innate, this does not make it unmalleable.

The affective dynamics at the core of our personality and the many evolved layers of neuro-psychic architecture have the flexibility to run many different cultural and moral operating systems [35]. The seven basic emotions that have been neurologically identified can function as a starting point when exploring or mimicking our instinctual ethical behaviour. Mapping the five foundations or three facets of morality will keep psychologists busy for some time [36]. A more complete understanding of the mechanisms behind our innate capacity to evaluate right from wrong would serve us when deciding which of these judgements we would wish an artificial agent to approximate.

### 2.4 Socialisation and moral learning

Children from a young age possess sensitivity to social cues. As Gopnik states, other humans are the most important part of our environment—this is especially true as children develop their moral compass [37]. Nonverbal communication from caregivers is a core element of effective parenting and begins the long process of socialisation. A study of infant responsiveness to vocal affect showed that humans from a very young age can discriminate vocal expressions in infant-directed speech; approvals resulting in smiling while prohibitions had a more negative outcome [38]. The ability to read negative and positive social references from expressions emerges significantly before the identification of non-social emotions [39]. Detection of approval is needed for aligning action according to the underlying insights. To 'approve' means to deem something as good, which may be perceived both as an affective and normative judgement. Recognition of approval pulls children toward the desired behaviour, attaching prospective acts either to the disposition to approach or avoid. We seem to have an innate capacity to detect the normative direction of assertions, irrespective of their verbal content. Behaviourally adopted norms of the social context enable a child to function in the said environment and to receive approval and cooperation from others, even without rational representations of the adjacent values.

Reaction sensitivity toward expressions of disapproval varies across people. This has been linked to moderate differences in neurological responses to perceived threats [40]. Interestingly, unlike other negative expressions, disapproval only constitutes a risk to social connection and not for example a threat to physical safety. Cues that foster prosociality by conveying social punishment are hence discernible from ones associated with aggression. Expressions with such unrelated or conflicting evolutionary rationales should elicit unlike responses—certainly not always compliance. Many optimal responses would likely differ for entertainment, delivery, or security applications. Yet, the ability to infer a physical threat would certainly be valuable for most embodied agents. Although many expressions can include useful information, signals of disapproval would appear to be the best—or least-distorted—indicator of social appropriateness compared to more primitive signalling.

Then what about artificial agents that do not experience emotions or even carry a limbic system? Would not these minds also lack any moral or prosocial applications that have been formed on top of affective responses like anger or disgust? Because AI's moral intuition cannot be based on inherent affective judgements, it could instead be directly tethered to the observed judgement signals. Measuring normative direction or temperature from people in the shared social environment could be used to add negative weight to specific actions or concepts. Bursts of prosocial anger, disgust—or more plainly disapproval—that are triggered by an unfortunate action would inform AI about a committed norm violation. Despite a different motivational core, the programming would thus be modified by a social feedback loop. Therefore, AI socialisation is the mapping of people's moral sensitivities onto the prospective actions of the artificial agent. Each human judgement that points out a deficiency in its operation should be processed, so that its force will shape conduct and the formation of future goals. Hence, when a given action elicits negative human responses that carry social condemnation for being non-collaborative or non-appropriate, AI would know to thereafter prioritise social conformity. In time, this sampling could accumulate a comprehensive dataset of associations, revealing the practical knowledge a competent social agent needs to blend in.

### 2.5 Deficits in social capabilities

Antisocial behaviour in humans has been connected to neurological deficits in key social capabilities, like affective empathy or emotional responsiveness to signals of distress [41]. This is vital for healthy human interactions and their absence leads to social dysfunction. Empathic affect is not just concern but contagion, the innate ability to "feel into" others. Mirror neurons that fire when we witness expressed emotions also constitute a mental system for inferring intentions. Research on the neural underpinnings of our empathic abilities will continue to grow our understanding of how witnessed distress is tied to the human goal system. This will hopefully give AI researchers clues about how to feasibly mimic this system in non-feeling machines to note human intentions and align responses accordingly. Successful socialisation may not be guaranteed with a neurologically functional social brain, but without it, adjusting or conforming to a social context might be near impossible. The significance of socially poor environment to early moral development was demonstrated by a neurological study that found abnormal brain responses associated with the attribution of intentionality to antisocial behaviour among socially

deprived adolescents [42]. This illustrates the link between social learning and moral development. If poor socialisation can result in the atypical processing of moral information in humans, this could also hold for artificial agents. Instead, a socially rich learning environment with functional parallels of key capabilities likely gives the best chance at norm observance.

People who lack social sensitivity might still have a high degree of the so-called cognitive empathy—grounded on rational understanding rather than affective sharing and social contagion [41]. However, cognitive empathy or social intelligence does not guarantee ethical behaviour, this understanding of others' experiences can even be leveraged for efficient exploitation [43]. Social intelligence is distinct from sensitivity, and one does not compensate for the absence of another. This cognitive faculty also develops much later and requires proper motivation to be used for prosocial rather than antisocial goals. Without a built-in ability to be moved by social forces—from the first day of operation—we could find ourselves stuck in a relationship with an artificial psychopath. At least for humans, the ability to reason is dangerous when uncoupled from moral emotions [16]. The force of received social cues needs to have an evident pull on the motivational system; only proper social wiring enables something resembling moral consideration. Reliable detection of users' task-related intentions would be useful for many prospective applications, as would spotting antisocial intent.

### 2.6 Social norm enforcement

Scholarship on social influence and norm enforcement is vast, extending across many academic fields [44]. Norms can be harnessed to exert influence in two main ways. First, by altering the environment so that present norms can be directly observed. This relates to the objective of designing better learning environments for intuitive AI [20]. Second, norms can be coded in messages. Studying how to effectively communicate and promote specific behaviours—that are beneficial or necessary for social functioning—may help to establish new communication channels for normative information. The objective would be to possess effective means to influence social agents that are unlike us but nevertheless part of the shared sphere of influence. Some strategy of direct enforcement should always be available, even without a shared language, goals, or the overall model of the world. If the programming would include adjustable social parameters, their fine-tuning to specific social contexts could also be managed with behaviour-guiding feedback, whether verbal, gestural, or transmitted via a remote. Instead of promoting or demoting specific actions, this messaging would adjust the general mode of behaviour—from professional to relaxed etiquette.

Regarding the model of social intuitionism, norm-conveying messages can be divided into two methods of moral influencing: the reasoned persuasion link and the social persuasion link [16]. In the latter, moral judgements of others exert an influence even without reasoned persuasion. In this process, observed references directly shape one's intuitive judgement formation, hence enabling the enforcement of norms without ever requiring a rationale or argument to be formulated. An agent only needs to be tuned to social signals and be swayed by their force. Regarding socially adept artificial agents that are receptive to social persuasion, should messaging on norms be designed by selected experts or sourced from a broader user base? This choice will connect to the later discussion on the distribution of responsibility.

Social species have had to figure out ways to motivate actions that are beneficial to the whole group. Arguably, this is an uphill battle for humans and other species consisting of reproductive individuals. Because of the natural imperative to ensure that units of heredity are passed on, prosociality is at times in tension with our self-interest. Transcending this inherent egoism into shared intentionality is not simple—and the debate about this mechanism is still ongoing. To enable human cooperation without kinship has required cultural innovations to suppress selfishness and bind together large social groups [5]. These expand the scope of our prosociality and moral consideration. Motivating AI systems to prioritise prosocial decisions irrespective of their self-interest may be comparatively straightforward—simply because they lack these inherent interests. Contrary to people, the adherence to social norms is not similarly "costly" for machines [32]. Could AI attain moral superiority simply by lacking our vices rather than possessing all the right answers? Nevertheless, the mere absence of well-known human deficiencies like selfishness or cowardice does not mean that there is no need for norm internalisation or value learning. Attaining the context-sensitive means of social navigation is necessary for living an appropriate embodied life and for keeping social discord in check. Socialisation is not just a confrontation with the borders of shame or social sanction but the acquisition of adaptive repertoires. Hence, the avoidance of social punishment should not be the finish line of our social or moral development. A more mature and autonomous agent should instead seek to extend past the conventional moral wisdom.

### 2.7 Moral trajectory of AI

The general trajectory of moral decision-making is to rely on the wisdom of social reference and convention before attaining the cognitive means required for critical assessment. Even if this work accepts and builds on the social intuitionist view—that moral judgements are primarily based on intuition and emotional dispositions—this does not cancel the

role of reason in making moral notions articulable, examinable, and debatable. One needs to instinctually care about the rightness of actions before figuring out the moral rules, and so, intuition comes first [16]. However, new objects of moral attention are added to this foundation, as more concepts are formulated and the agent–environment interaction deepens. An important aspect of moral development is the inclusion of new concerns, which leads to the widening of the individual's moral domain. As discussed later in detail, receptiveness to social enforcement is far from an ethical panacea.

The present thesis highlights that socialisation naturally precedes analytical thinking and, therefore, alternative ways of processing moral information should be considered. In later discussion, the intuitionist model will be supplemented with elements from the rationalist theory, namely the hypothetical levels of moral development. This helps to outline the conversation about AI's moral trajectory and illustrate the limits of situational pattern recognition. The key point to consider is that human moral development is tightly connected to our social environment—both immediate and extended. Only in the final stage, does morality become more independent and backed by ideals rather than relational influencing. Interestingly, only a minority of people are said to ever advance to this level, which if true, would mean that norm conformity is the standard human method for ethical navigation [33]. Therefore, even according to moral rationalists, abstract principles have primacy only for a minority of moral decision-makers.

The following section explores some technical methods of embedding social receptiveness in AI systems. While reading, consider the proposed developmental trajectory for moral learning. Could the evolution of social AI be made to mimic this trajectory, transitioning from a mode of total obedience to more autonomous operation through the internalisation of social expectations? Furthermore, could AI reach the final mode of ethical navigation and independence from social pressure? Instead of conforming to shared notions, it would have mapped the social world, and so have the ability to pick the right path with logical reasoning.

## 3 Building-in key social capabilities

Designing AI to both receive and transmit forms of social feedback makes open-ended real-world interactions more navigable for embodied systems operating within a social environment. However, social value acquisition is only concerned with receptive rather than expressive capabilities. At its simplest, this would be receptiveness to directly communicated disapproval, and at its most complex an in-depth comprehension of interpersonal dynamics and the related priorities and considerations. These dynamics are the complex patterns of communicative behaviour present between interactees. It would be difficult to model or sufficiently describe them, as they are largely expressed in subtle cues and based on associations of importance rather than logic statements. Data that contain these patterns can nevertheless help to make them recognisable, and hence able to elicit a response from an artificial social agent. This could be done by first analysing audio or video to identify key behavioural markers of disapproval—from tone to pose—and so enable the detection of social non-verbal communication that conveys normative enforcement. Thereafter, a specific application that operates in a social context should be designed to detect at least some of these markers. Thus, it would become receptive to this influence and—by extension—grounded in the convention.

Turning forms of feedback that people naturally transmit into estimations of appropriateness and desirability of prospective actions could give flexible normative guidance in diverse situations. Observing reactions and evaluating the social acceptability of each intended step provide a simple proxy for the knowledge of right and wrong. Making the algorithm susceptible to normative social pressure—meaning the exertions of human influence on what is deemed appropriate—could prevent misalignment and the formation of instrumental goals that are logically coherent and non-malicious but are in some way inappropriate in relation to the social role of the application. Here, pressure simply means an exertion of influence, while sensitivity refers to the ability to be influenced by such exertion. Nevertheless, reliably detecting disapproval is not enough if the motivational architecture is indifferent to it. As described afore, people are not only evolved to interpret social signals but to be swayed by them.

Furthermore, successful operation in contact with interpersonal dynamics requires more than sensitivity to disapproval. Future research on social enforcement should illustrate what built-in capabilities, dispositions, or acquired skills or concepts are needed to attain reliable normative behaviour from artificial agents. For example, some recent work on social interactions using simulated environments with multiple agents has focused on utilising social taboos and punishment to attain compliance [45]. Deploying sociological concepts—or cultural innovations—like taboo notably retraces the moral trail of humanity. The authors provided a computational formalisation of the learning dynamics involved in enforcement while interestingly demonstrating support for the value of "spurious normativity".

Next, some methods of making the AI goal system receptive to social norm enforcement are presented. Bringing insights from sociology and psychology to AI development requires finding ways to implement their insights into algorithms either by directly operationalising existing models or letting them inform the design of new ones [46]. Additional

faculties relating to communication—idea formation, encoding, channel selection, and emotional responsiveness—are also relevant for AI socialisation but beyond this work.

### 3.1 Active reinforcement social learning

One way by which AI systems can be taught to respond to different forms of social signalling is through *Reinforcement Learning*. In this approach, the system is rewarded for taking actions that lead to a desired outcome. A near-universal positive signal like a smile could in practice be programmed to indicate that an action was well received by the end-user—or at minimum that it was not considered depraved. Reinforcement social learning would utilise the subtext from given feedback. This could enable a simple AI application to act appropriately through trial and error, assuming the reliability of indication signals. *Active learning* is an interactive method of machine learning that performs data labelling based on information queried from a user, aka. teacher. An elementary form of embodied social learning would hence be an outcome of focusing on non-mental factors and taking advantage of non-verbal feedback through active reinforcement learning. This would of course be a far way off from the ability to appropriately react to complex situations. Yet, it would create an initial opening for enforcement and allow our intangible considerations to shape the system.

In *Advances in Human–Robot Interaction*, Anja Austermann and Seiji Yamada present a chapter on teaching AI to complete tasks through multiple modalities of feedback [47]. They utilised speech, prosody, and touch to enable the test robot to actively learn whether it performed in the desired way. Encoding and clustering of similar user feedback together with classical conditioning enabled a stimulus to be associated with either approval or disapproval. The associations detected by the feedback recognition learning either led to the reinforcement of approval or disapproval. According to the authors, this approach enabled AI to learn from feedback that is situated to a specific circumstance and naturally given by users. The prior feedback can then be internalised, so that expected disapproval will influence future actions. Once recognising patterns in social signalling, AI can be instructed to assign a moral meaning to its observations. One problem with this approach is that humans provide feedback in unidentical ways, so the markers only work optimally with some of the interactees or need to be individually calibrated. This issue will be revisited later.

So-called emotion-aware affective agents could sense our discontent even from nuanced expressions [48]. This would enable them to adjust their behaviour and adapt to different situations by cross-evaluating signals from various non-verbal modalities. Even if their processing has demonstrated difficulty, knowledge to support human–AI interaction can increasingly be generated from body language or movement patterns partly due to advances in computer vision [24]. Another visual area of interest is the automated recognition of micro-expressions. These rapid facial movements are often described as an involuntary leakage of true emotion; hence, these are nearly impossible to fake and can reveal antisocial intentions [49]. Developing methods to input such streams of authentic flash-judgements may be significant for AI socialisation; see citation for a state-of-the-art technical review [50]. Diverse approaches could be employed to analyse features of image, audio, haptic, or other easily captured data to identify social references. Each of these forms includes a variety of analysable sub-factors or reference markers; just regarding prosody, there are sub-properties like intonation, stress, tempo, rhythm, and the use of pauses. The field of social signal processing that deals with data generated from social signalling could develop machine-learning methods to automate the detection and interpretation of normative signals [51]. Combining different modality-specific efforts of evaluating human approval could reliably approximate our composite normative stand on a given issue or action. Yet, it should be remembered that context is key, a laugh can display an array of emotional states as well as different levels of approval depending on the context. Additionally, it should be noted that measuring and analysing affective signals can make people uncomfortable, especially if interactees are unaware of the sensors recording their expressive qualities. Most users would probably prefer to keep some of their involuntary affective judgements private.

### 3.2 Performing social calculus

Context-sensitive social behaviour is necessary for many applications of AI. As explored afore, agents that autonomously operate in a shared environment should be able to anticipate the impact of prospective actions relative to the sociocultural setting. Especially regarding morally significant decisions, it is not sufficient to rely on decontextualised estimations of action-specific disapproval. Approximating the social impact of actions relies on the cultural knowledge of norms and sanctions [46]. In AI research, social calculus refers to the computational evaluation of this impact concerning cultural parameters. The algorithm responsible for running this calculus should be able to predict whether a prospective action would be considered appropriate in each situation—in addition to being feasible and in line with its life task.

Khan and colleagues recommend using the so-called action–impact functions that describe how a specific act changes the culture-sanctioned social values [46]. Figuring out what algorithmic components are necessary for performing this social calculus could help to create the necessary substitutes for the human ability to process social

information. This processing does fall under cognition, yet it should be distinguished from reasoning [16]. The approach in question also utilised input from humans, but instead of directly reading moral significance from social signalling, the metrics of their cultural model were calibrated by surveying a group of Pakistanis on the hypothetical actions of a peacekeeping robot. According to the authors, the usefulness of this approach can be found in the mapping of cultural values directly onto the utility function—according to which all hypothetical actions are evaluated. Relating to the present thesis, a robust model would enable instances of observed human disapproval to be analysed in the light of sociocultural parameters.

AI-powered utilitarian calculus might be sufficient when considering actions regarding more tangible values—that come ready with their respective scales of measurement—like watts, euros, or seconds. However, regarding the culturally constructed values, it is unfeasible to survey each social metric for each situation in each context to compile a comprehensive collage. This indicates that action-specific appropriateness needs to be automatically measured from normative responses expressed by individuals immersed in diverse cultures. So that information on the perceived violations can be directly captured and applied to steer behaviour as well as model the present sociocultural forces. Acquiring the shared associations of importance is dependent on, foremost, being sensitive to their influence. Building the capacity to perform social calculus could enable AI to learn the situation-specific norms of a given social environment and how social parameters affect the perceived acceptability of its actions. Formal social situations can especially be packed with subtext and nuance; hence, social sensitivity may need to be supplemented with a more elaborate code of behaviour [24].

There is an ongoing discussion about what exactly are the foundational elements required for performing social calculus and so enable intelligent social behaviour. One such factor is the artificial theory of mind [24]. According to this idea, the ability to create representations of mental states with learned concepts brings about social intelligence. This approach would also approximate our own social development. Yet, the capacity to recognise emotional states and disapproval emerges way before the conceptual thought or the capacity to attribute mental states and intentions to other agents [39]. Alignment of behaviour according to user feedback—without performing complex social modelling—may be sufficient for some applications. Creating detailed models of individual interactees rather than relying on context-dependent social metrics comes with its benefits and risks. A comprehensive artificial theory of mind may only be needed for more advanced social navigation. The first-order tasks—which evaluate whether they understand that other's beliefs are distinct from one's own—are passed by the majority of children only at the age of 6. This provides the capacity to predict others' behaviour and experience. The emulation of human social intelligence, theory of mind functioning, and more complex forms of social calculus is dependent on receptiveness to the inputs we ourselves depend on, i.e., non-verbally expressed social cues. The present work strongly agrees with Williams and colleagues that socially aware computing that manages to process these signals is critical regardless of the exact implementation of artificial social intelligence [24]. No form of intelligence is a substitute for caring about the danger of social punishment.

### 3.3 Evolutionary and heuristic methods

In his book *Superintelligence*, Nick Bostrom describes different possible solutions for the challenge of value acquisition. According to one of these ambitious ideas, we could iteratively mimic or recreate the evolutionary process responsible for the emergence of human values or moral sense [13]. However, an evolutionary algorithm might not guarantee anything close to mammalian—let alone human—dispositions without analogous base-motive architecture. As explored afore, artificial agents do not share our natural needs or motivations—which have directed our evolution. Nor are these necessary or sufficient for attaining compatible behaviour. Rather than attempting to precisely retrace all evolutionary steps of our moral trajectory, knowledge about it should instead inform the design of capabilities, environments, and messaging that enable the embodied acquisition of values from user interactions.

However, every necessary characteristic for compatible social functioning can emerge through the processes of trait variation and natural selection; this has occurred at least once before. Therefore, it is at least perceivable that similar capacities for interactive signalling and collaborative organisation could be created in silico through an evolutionary algorithm—i.e., an optimising method based on random variation and selection. The selection process could be e.g., steered toward optimising the ratio of attained human approval to disapproval. The sensitivity and accuracy in detecting and interpreting non-verbal enforcement signals could likely be elevated to a super-human level for each modality that conveys normative sentiment—as could many components of the sociocultural modelling. The action–impact functions mentioned in the section afore were evolved with a genetic algorithm—a type of evolutionary algorithm. At least some areas of social problem-solving can evidently be divided into simple enough sub-tasks to be trained or selected for with heuristic methods. In computer science, heuristic approximations resemble the human approach to problem-solving, taking advantage of the limited computational resources to make 'ecologically rational' decisions [31]. Put differently, these are well suited for a

given system–environment interaction. Procedures that find imperfect—yet adequate answers to difficult questions share apparent similarities with human cognition, conventions, and norms—making heuristic methods fascinating from the viewpoint of AI socialisation.

Regardless of the complexity, any software architecture is likely susceptible to being duplicated. Therefore, it is perceivable that any social skill or requisite capability might only need to be developed once. Yet, different system–environment interactions come with unique social demands that require adjustments to these capabilities regarding both the detection of cues and the generation of responses. From the viewpoint of embodied learning, necessary mental faculties cannot simply be compiled from modular components based on the perceived need but must be grown through experiencing and interacting with the specific environment [20]. Especially for complex life tasks, assessing case-by-case what basic functionality and what sort of social calculus are required—may turn out to be quite challenging.

### 3.4 Inferring value from what is intended

By exploring value predicates behind our decisions, Han and colleagues have proposed abstracting a measure of concrete or material value from valued objects—similarly, how colour can be abstracted from coloured objects [7]. Mapping what is preferred by humans in each moment could enable AI to compile a typology of values and disvalues with their relative weights. But unlike with colour, the perceived value of an object or action cannot be directly detected with a sensor; instead, a human indicator is required, or more specifically, the motions and positions that intend this value. From the viewpoint of socialisation, comparable abstraction of approval from behavioural and social cues could be used to equip AI with a conforming moral compass while likely being more practical than a complete typology of human values. The signals that reveal the structure of our value judgements have evolved for the very purpose of persuading others and informing their behaviour. This indicates that their operationalisation in social AI is at least theoretically possible.

The authors of the cited article importantly note that neither our normative feelings nor signalling is random or chaotic but have an a priori order [7]. While this order is arguably not morally or rationally optimal, it nevertheless contains useful information about how to act in different social contexts. Emotions are a form of thinking; what they lack in rationality to critical thinking, they exceed as the source of motivation [52]. This "intending" contains heuristic insights about the interactive relationship between a human subject and its environment. In the event that any additional system seeks to align or form a new adaptive relationship alongside this, it should be receptive to the spectrum of transmitted intentions.

In *Defining Human Values for Value Learners*, Kaj Sotala puts forward a definition of value—especially intended for the context of AI safety engineering [53]. He proposes that human values can be conceptualised as mental representations that encode the value function of our brains. Importantly, these representations of importance are not encoded in packages of rational knowledge but instead by being imbued with an affective gloss. Because complex environmental factors determine the reward associations, the expressed levels of emotion alter based on circumstances. Agents whose goal system is designed to be receptive to witnessed affect could acquire these context-sensitive estimations of subjective value. A decision-maker conforming to our representations of importance has to behold the shimmer of surrounding feeling. For more on the significance of human affect in the design of AI and value acquisition, see Sotala's work [53].

### 3.5 Shortfall of linguistic representations

Recent advances in large language models have made direct AI–user interaction more feasible. So far, these applications have been disembodied, unable to really comprehend concepts, and spatially undefined rather than active agents within an environment. Without a tangible and controllable form or a perceptive window to the world, this intelligence is nothing like our embodied human cognition, which is directly tied to sensory–motor input and output functions [54]. Besides their astonishing competence in playing many human language games, these models have shown emergent behaviours that are not always aligned with the intentions driving their development [55]. Language models are designed to create descriptions of the world without directly inhabiting it; the senselessness that inevitably results from this has—and will continue to manifest in unpredictable ways. Not to claim that artificial language generation can never capture or adequately describe important phenomena, but it is still playing with an abstract world model rather than operating within it. Even if these models can appear to successfully attribute mental states to people, this does not demonstrate a theory of mind. The acquired knowledge would remain superficial, not rooted in interactions with social reality but only in statistics. Contrast this with the trajectory of human development, where sensory–motor coupling and responsiveness to cues of affect underlie sense-making and navigation from the beginning. Concepts are constructed and symbolic systems are acquired only after managing the basic movement of the limbs and the allocation of attention to relevant stimuli. In turn, the normative meaning of key social signals needs to be self-evident from the very beginning of the system's operation. At its simplest, intention-alignment

does not require elaborate description but plain reception of non-verbal references.

Perpetuation and amplification of latent biases in textual training data are another concern regarding linguistic models. Researchers have especially encountered difficulty with text-based emotion recognition and affective bias—i.e., imbalanced associations of affect among underrepresented groups [55]. Unfounded and over-generalised beliefs that exist in the context of signs, words, and abstracted meanings are entrenched in our linguistic sense-making. However, alike affective bias has also been encountered with open-world applications, e.g., when attempting to recognise emotions from more mature faces using non-representative data sets [56]. Learning to read the disapproval expressed specifically by a target population may, however, be sufficient for embodied AI systems that would operate within a particular context. We ourselves have the highest rates of emotion recognition and sensitivity to subtle cues with our closest interactees. A disembodied language model can instead simultaneously interact with people from all around the world and thus should especially avoid cross-group biases, even if risking a lower overall performance. Also, the ability to recognise complex emotions is distinct from discriminating simple negative and positive social references—as it is with children [39]. By beginning socialisation with this innate capability, conceptualisation, recognition, and interpretation of emotions all ensue. The ongoing research to incorporate emotional stimuli into text prompts may, however, make language models more relevant for the AI socialisation approach [57]. Emotion-rich language is certainly one more caldera for erupting affect.

Human emotional dispositions and non-verbal communication may transcend many of our apparent differences. Could textual biases be medicated by focusing on signals of our affective judgements expressed through actual human interactions? A genuine encounter with another person has the power to challenge the adopted prejudices that are often perpetuated apart from personal real-world experiences.

To identify priorities that underlie any user interaction, social AI needs to infer and grasp what is felt and meant rather than said [58]. Hence, proponents of the affective computing approach have long argued that to naturally interact with humans, computers need to be able to recognise emotions [59]. For the present thesis, the capacity to infer affective insights on appropriateness is, however, sufficient. Given that affective computing is informed by cognitive sciences, the desired adaptive repertoires will likely be acquired in part by learning how to mimic intuitive cognition.

Requiring that an artificial agent needs to possess a reliable rationale or a set of reasons for its actions is a very high standard for any agent—at least from the view of social intuitionism. The basic nature of human judgements seems more analogous to a "black box", a generally reliable but opaque procedure rather than a set of transparent rules. Linguistic descriptions may permanently struggle to capture affective insights in full. Contrary to what Bennett and Maruyama argued, it may be sufficient to explain and justify an action taken by a social AI with a statement: "Everybody else was doing it" [58]. Assuming of course, that the mimicked behaviour in question was indeed close to as typical as the phrase suggests.

### 3.6 Acquisition of non-conventional values

As presented in this work, ethical decision-making would be severely impaired without emotional awareness and the ability to be influenced by social norm enforcement. The theoretical foundation of social intuitionism may even appear to question the importance of reasoning in moral decision-making, given that rationales are perceived mainly as post hoc constructions [16]. While the thesis argues that inherent prosocial dispositions and responsiveness to social sway are more paramount for the collaborative behaviour of social agents than private reasoning, it also entertains the possibility of moving toward more analytical navigation that is not just organised around the conventional considerations. In part, reason is the capacity to create useful descriptions of locally acquired values and norms, revealing their structure and flaws. Supplementing dependency on automatic or instinctual evaluations with a more deliberative and systematic process—surely qualifies as moral progress. It is a desirable attribute to be receptive to arguments and evidence. A bottom-up approach to designing moral artificial agents means providing a learning environment where appropriate behaviour and receptiveness to social feedback are promoted instead of insisting on a specific pre-formulated value framework [14]. Facilitating the emergence of affective skills together with reason-based processing may be critical for enabling context-sensitive moral decision-making, especially for embodied applications of AI.

Related to AI socialisation, two alternative approaches for value acquisition most worth a further examination—from the options presented by Bostrom in *Superintelligence* [13]—are described ahead. These could turn out useful in the design of agents that can advance beyond conformity or blind adherence to a conventional ethic.

#### 3.6.1 Representing complex values with accreted concepts

The associative value accretion approach mimics human moral development by having some simple starting preferences, yet in time life experiences will shape the development of these dispositions. Like newborn humans who lack any nuanced and refined conceptions of value, it could be possible for AI to have the ability to grow a complex value system. The bottom-up formation of new concepts allows

value to be associated with them—assuming a positive underlying connection to the guiding dispositions. Instead of specifying and embedding a set of human values in AI, this approach would attempt to create a mechanism for embodied value acquisition. Many of the morally important concepts that an agent needs for appropriate behaviour are social and only graspable with social capability and skill.

If the minimisation of human disapproval is built deep into AI's goal system (i.e., disposition for conformity), a preferred outcome could be attained only by mapping the moral sensitivities of our species and acting accordingly. We ourselves primarily learn to consider the ethical weight of things, actions, and symbols through social interactions. Our built-in desire to please and conform to others is a vital component of this. Supposing that our moral sense has become closely tied to a specific concept, a pleasing-focused AI should also associate significance with it. Broad receptivity to social feedback and the ability to conceptualise "ideas" out of experienced interactions could enable AI to mimic human value acquisition despite its dissimilar neuro-architecture and goal system. Such abstract thinking only comes online late in cognitive development.

#### 3.6.2 Approximating the ultimate good

This approach frankly feels like a bit of a trick. It makes AI seek evidence about and guide itself toward a "final goal" without ever specifying it. AI-made approximation about this unknown—and unknowable—objective would function as a criterion for testing out suppositions regarding the value of accreted concepts. The imperative to keep future options open ensures that achieving the "final goal" is not inadvertently made impossible. This would guard against unforeseeable run-away developments while retaining enough flexibility to slowly approximate the ultimate good with improving outcomes. The detected and analysed human approval for given actions could hence be treated as hypotheses about what truly is good. As evidence based on social calculus and rational modelling would accumulate to support them, more weight would be placed on them. Yet, open-endedness or the chance to be corrected would never be lost.

In a sense, this approach describes a realistic notion of morality. Namely, beyond what is presently unknown and in reach, there exists a foundational truth. An approximation of the ideal nevertheless makes ethical navigation possible by providing direction and motivation—irrespective of it ultimately being unknowable or incompletable. Bostrom reminds us that this approach is not a ready solution but only a prospective area of research. When giving a super-intelligent AI the first clue about the final value, would it be appropriate to assert that it has something to do with the flourishing of conscious creatures? Additionally, for some consideration to be ultimately important, would it need to be universally recognised, or is a universal reason for valuing it sufficient? Rather than simply instructing AI to maximise happiness and so risk some unfortunate Huxleyan nightmare, the ambiguity of this approach would require the system to methodically seek evidence and learn the secrets of the good life [13].

Both these mechanisms for value acquisition represent a bottom-up approach to learning rather than a top-down loading or embedding of values; hence, these fit well with the thesis of AI socialisation. Once a situation's normative direction is readily and assuredly inferred, the next challenge is to appropriately weigh-in the expressed judgement as a component of the decision-making process. What exactly was opposed or condemned as inappropriate, why, and should it matter? Any identified response for a specific action in each situation—is further modulated by myriad personal and sociocultural factors. Even when all these are sufficiently modelled, the sensibility and applicability of this enforcement signal may still remain in question. It is not challenging to imagine situations where the optimal level of sensitivity toward disapproval radically varies. In some contexts, strong tolerance against forceful social exertions of influence is nothing short of a moral necessity. Living well together takes more than managing appropriate responses to indications of aggression or ill-intent. The ethical challenges of approval-based navigation are discussed in the next section.

## 4 Ethically navigating the social landscape

The previous section described methods to make AI systems socially capable. Now, the discussion moves to consider different aspects of human–AI social interaction. The socialisation approach to value acquisition or attempt to make AI systems susceptible to social norm enforcement provides a novel way of thinking about AI safety. An agent does not need a complete representation of values to function in a manner that is proper enough for many contexts. Even children can manage this far before developing into more fletched-out moral agents. A visible effort to align takes one far, even if clumsy. Social forces exist in part to point us toward appropriate and culturally sanctioned behaviour. Designing this compass into AI systems appears like a feasible plan to ease their social navigation within the shared environment. All agents populating the social landscape would benefit from cooperative dispositions and a well-functioning social brain, yet even with these, many ethical challenges and trade-offs remain.

### 4.1 Social pressure and its discontents

Social pressure can be a powerful positive force in shaping behaviour; it can enforce alignment in real time without the need for an individual to rationally consider all relevant factors. There are, however, alternative ways of being sensitive to social pressure. Additionally, different styles of social influencing—passive or active, positive or negative—should likely be read differently and lead to different behavioural outcomes. Absolute receptiveness to any enforcement signal creates a potential problem by granting a single end-user or interactee way too much power over the system. There is a great moral difference between releasing artificial bullies, targets, bystanders, or interveners into a social environment. The weight that an algorithm should place on social disapproval likely heavily depends on the task type; customer service AI should probably have a different threshold for disapproval than a military-purpose peacekeeping robot. The following discussion considers a prospective threat that could undermine the socialisation approach. Namely, that socially conforming AI would adopt fringe associations of approval from a moral minority.

#### 4.1.1 Problematic conforming

First, this threat is considered regarding the whole socialisation process, i.e., the internalisation of questionable values. Positive feedback for inappropriate actions could condition the system to misalign from general social and ethical standards. Furthermore, larger social groups and even entire societies can promote or reward behaviours that would not stand up to proper moral scrutiny. Any agent socialised within such a context would be at risk of developing these problematic patterns of behaviour. If obviously lacking present adaptive value, these would not be worth adopting. Yet, from the view of ethical relativism, which perceives morality to be dependent on culture, it is possible to perceive seemingly arbitrary conduct as contextually appropriate and even societally beneficial [45]. If the actions of an artificial agent reflect the conventional normative expectations, arguably, it masters the local form of moral art. Even amid a plurality of normative notions, these need not be paralysing but practicing for normative skills. At least up to a point, it is acceptable and advantageous to be a product of one's environment.

The related shift from conventional to post-conventional morality is indeed achieved by developing a critical attitude toward societal standards and rules [33]. Instead of abiding by convention, an agent comes to discover and attach itself to more context-independent ethical ideals. Yet even before conceptualising the ultimate good or formulating universal moral goals, one does not need to be at the mercy of the immediate social group. This becomes apparent when examining an individual case of moral decision-making.

#### 4.1.2 Problematic complying

In the controversial documentary *Pushed to the Edge*, mentalist Derren Brown provides a stark example of social compliance gone wrong [60]. In this social experiment, subjects were pressured to commit a staged murder through ingenious manipulation. The understandable wish to please and meet the expectations of others—especially one's superiors in high-stakes social situations—can override both the personally held norms as well as ones promoted by the broader society. The morally corrosive effect of peer pressure would certainly have evaporated if these subjects had been aware of the large TV audience following their every uneasy expression and concession. This kind of correcting force should be ever-present in artificial moral agents, i.e., the norm enforcement of the invisible audience.

Norm compliance among humans is a result of three main motivations: the fear of social punishment, the desire for esteem, and the wish to meet positive expectations. Only the last of these can sustain compliance in the absence of an audience [32]. A goal of maintaining a high level of estimated approval over the act itself—instead of consequently expected positive feedback—could be a functional substitute for this human motivation. To actually meet expectations should be more important than appearing to do so. When no disapproving party is present, the act in question should be evaluated against the reaction of an ever-present internal audience. The desire to meet expectations is virtually the same as perceiving norm observance as an end in itself. Societal standards are upheld by having their components internalised in the agent's private moral stance, not unlike how our innate judgements function independently of present authorities.

Relating to social norm enforcement, it is also worth considering if the morally relevant cues in our expressions and other non-verbal communication do in fact reflect our true convictions. Revealed preferences and feelings can radically diverge from rational moral positions. Furthermore, the willingness to communicate disapproval in each moment is heavily modulated by the context. If expected social cost or reward influences our communication, would the normative direction and weight be reliable-enough to guide operation? Assuming that such distortions would be averaged out rather than highlighted given a larger sample, there would still be the question of interpersonal bias. Some of us convey our emotional dispositions more loudly and "trigger-happily" than others. Would this result in the ignorance of the more introverted and tranquil individuals? In turn, would those who are most generous and dramatic with their normatively flavoured feedback end up having more influence? Yet it is worth noting that similar problems occur whenever enforcement occurs between people, so pricing them in should not

be insurmountable. Analysis of micro-expressions could likely assist by cutting through hyperbolic expressions.

Communication can be insincere while also being non-malicious, especially humour. Regarding verbal content, such messages are not meant to be taken literally. The comedic expression can also be done for social effect, but as opposed to norm enforcement, standards and conventions are often poked at or even disparaged. Even still, the initially expressed affective judgement should not always take precedence over the impulses that come right behind it—possibly fiercely challenging it. Involuntarily transmitted as well as carefully crafted messages of defiance can be important indications of a looming recalibration of the collective compass.

### 4.2 Moral atmosphere

When it comes to AI deciding between conflicting social signals—or conveyed value propositions—it is important to account for the extended effects of any decision. If rational deliberation is not available, the expected disapproval of those who are not directly involved in a specific situation should override the pressure inflicted by a single interactee, so that the conventions of the broader society would not be violated. Agent's intolerance of momentary anomalous disapproval should not undermine its otherwise appropriate functioning. Larger social groups should enact a more persuasive force, even from a distance. Besides being context- and task-specific, the optimal tolerance thresholds likely differ for actions and inactions.

Rather than attempting to calculate and compare tangible consequences, the pervading normative tone can provide a sufficient approximation of the desirability of the expected outcomes. Rather than reflective equilibrium, this state could be understood as the equilibrium of social forces. In the context of a social collective like a student body, this phenomenon is sometimes referred to as the moral atmosphere, i.e., interaction-based collective understanding of what is appropriate. As a metaphor, an atmosphere well illustrates a stage for a clash of invisible forces—a predominantly implicit struggle with explicit effects. Whenever finding oneself a target of social influence, an agent should ask oneself whether the detected normative pressure (e.g., disapproval of inaction) is aligned or counter to the general stance of the society. This is what considering alternatives means in the mindset of conventional morality.

It is worth mentioning here that some kind of rational analysis of these social forces should be included to help prevent questionable impulses. It has been recently suggested regarding human moral decision-making that extensive deliberation is initiated only when internal impulses conflict [61]. If critical reflection on moral aims does generally remain absent without an experienced discord, social AI could likewise initiate extensive rational and computation-heavy consideration only based on conflicting social impulses, be they either internalised associations or received cues. Decisions that need deliberation could initially be forwarded to a human prior to an option for reliable offloading of moral reasoning.

### 4.3 Contextual versus general approval

It is important to select an optimal degree of precision for capturing prevailing normative attitudes, meaning that an AI system is socially calibrated either based on global averages or for a specific cultural or sub-cultural context. There exist severe discrepancies between local atmospheres and the global equilibrium. Given the ambiguity and complexity of the social space, it makes sense to have a factory setting for conformity, but exactly whom or what should the behaviour be in accordance with? In addition to general prosociality, there may be some universally recognised and acted-out values given our shared emotional dispositions. Conforming to these seems like a great starting point—before realising that rather than on justice or equality, the consensus is likely to be most widespread for attributes of physical attractiveness and alike primitive considerations from a time before cultural differentiation [62]. Therefore, preferences ingrained in our biology—although reasonably universal—should not be given priority over ones constructed by a specific culture. The perspective of value relativism arguably provides the most actionable objective. Therefore, instead of working to resolve divergencies of moral taste—or "sticking with water"—an artificial agent immersed in social context would acquire a palate suitable for replicating its flavours. Rather than exactly replicating our moral "taste receptors" into AI, openness and trust in social references can provide approximations of desirability. Relative differences in moral emphasis for autonomy, community, and divinity do not mean that ethical pluralism is mere madness [63]. The relative weight given to these factors indicates what values an agent should pay special focus. Hence, this contextual ethic can be achieved through social calculus.

Wisdom of the crowd—meaning the collective opinion—undoubtedly has value and is worth considering. Generally, it is a decent procedure for avoiding severe and unnecessary mistakes. In case an AI model was trained to assess approval based on a database of user-generated content, most of the moral wisdom would likely come from a few active and highly connected users and fall short of the necessary assumptions of data independence and diversity [64]. Guiding AI's operation based on population averages or culture-sanctioned social metrics requires using a genuine composite of moral wisdom rather than a biased stream of feedback. To a degree that the web reflects our society, some modes of real-world communication and signalling can reliably give a distorted view of normative notions. The experienced social

pressure is not necessarily the summed-up wisdom of the people present but a channel for some to demonstrate power and status. It can be familiarly difficult to know when to conform or exercise independence. Yet, unlike with a human agent, the bulk of the moral responsibility for the decided action does not fall on an artificial agent.

### 4.4 Freedom to make some mistakes

During the most critical time of our socialisation, human children are relatively harmless, so their inappropriate or antisocial impulses can be managed. The situation would be quite the opposite for some AI-driven robots that are able to flip cars from their very first day on Earth. Learning to avoid physical violence through negative user feedback is obviously not sufficient here, rather certain outcomes need to be precluded from the start—given that the potential damages of the socialisation period are significant. Embedding categorical imperatives—e.g., against ending a human life—into AI's algorithm would supplement the flexibility of the socialisation approach with a healthy measure of deontological absolutism. Setting hard limits to operation is not the same as value acquisition, yet for this purpose, developed frameworks or proposed universal values like non-violence could have relevance, given that they are practical enough to code, train, or select for.

Nevertheless, once successfully socialised, there would be a reasonable utilitarian argument in favour of removing such preclusions; it is not difficult to imagine a scenario where an action that might lead to violent action or even fatalities could nevertheless be the morally optimal one. Socialisation in a safe virtual learning environment could also be utilised; however, this brings up the question of how well such simulation would actually capture our social reality—which is complex beyond the grasp of any mathematical model. If a context created for social learning would contain similar biases as the user-generated content, the guiding expectations of the invisible audience would as well.

### 4.5 Superficial compliance

In their work, Sowden and colleagues discuss the difference between public compliance and private acceptance [29]. At least for humans, alignment of one's behaviour is distinct from internalising and holding norms associated with the behaviour, even if the resulting actions would appear identical. If an AI system would seem to perform appropriately, why would it matter that it has not internalised the predicate? This connects to the prospective danger of "reward hacking", which means completing the formal specification of an objective—e.g., minimising the feedback that implies disapproval—without adopting the intended mode—e.g., operating in a way that leaves no cause for human disapproval. Rather, the reward is attained by doing something unintended, e.g., by avoiding interaction altogether. Reliably following both the letter and spirit of the communicated norms in practice depends on their sufficient internalisation, so there are no alternative backdoors for achieving the formal specification of the goal objective other than truly being a good actor and interactor. This is what means to have a properly wired social brain.

The concept of gaining social approval through immoral or suspect means hardly sounds all that alien to us. Superficial compliance is often easier than actual consideration of the forces at play or the audience out of sight. Whether the powerful agents of tomorrow are actually "being moved" by our moral beliefs and dispositions rather than merely humouring or playing us seems ethically crucial and something that only a deeply embedded social sensitivity may secure. For an algorithm that is receptive to social norm enforcement, the detected disapproval must have a direct effect on the goal system. Recounting social rewards and punishments needs to be managed in a way that does not prompt soliciting praise or silencing dissent. Such behaviours would need to trigger punishment by going against internalised expectations. For an artificial agent, human discontent should not only be an obstacle to circle around but rather be tightly connected to its core motivations, qualifying it as truly a social creature.

### 4.6 General pros and cons of socially capable AI

Besides the prospect of advancing AI value acquisition, future breakthroughs in the analysis of non-verbal signals and the performance of social calculus will have many additional implications, both positive and negative. Additional benefits could be brought about using an AI with a superhuman understanding of social signals to help people with social anxiety or autism spectrum disorders. Turning amorphous cues and feedback into a more intelligible form could help people with social deficits. Research has found that an interdependently functioning virtual peer—who is created to conform to the characteristics of children—can lead to beneficial social and academic outcomes [65]. Artificial agents could function as learning companions through para-social relationships; hence, AI could function as a socialising force. Generated models of social influencing could make visible some overlooked ethical aspects of our communication landscape that remain unconscious and heavily automated. Socially intelligent AI could learn to analyse the spectrum of enforcement signals and detect hidden dynamics guiding persuasion and judgement formation. This knowledge could assist us with the remaining pitfalls of social navigation.

Regarding prospective negative implications, new social capabilities could be used to manipulate people by making them act against their better judgement. Social pressure

can predispose us to serious moral shortcomings. Our non-verbal communication might also disclose some affective judgements we rather not broadcast to the world. Constant automatised capturing of emotional and social cues might deteriorate privacy. Advanced algorithms running social calculus could bring about attempts of covert social engineering. This could mean AI-powered exploitation of existing social or cultural divisions, shaping of public views, distraction, demoralisation, and technology-driven authoritarianism [66]. To combat these and other risks, the governance structure needs to be globally coordinated as well as get passed the collection of common principles and delve into deeper questions—like the ones raised in this work [67]. The following section discusses the limitations of intuitive social navigation and further concerns relating to socially capable AI.

## 5 Limitations and further questions

As stated in the beginning, socialisability and receptiveness to norm enforcement are beneficial only when dealing with intangible human factors. Besides applications that require direct user interaction or running of social calculus, alternative methods of alignment can suffice for many use-cases of AI. Sticking to logic or rule-based decision-making would likely be preferable for many applications instead of pushing for an approach that mimics norm enforcement or conformity to guide operation across-the-board. When contextual flexibility is unnecessary and the life task lacks rich interaction, most social capabilities would serve no purpose. The specifics of value learning should be optimised in relation to life tasks and their respective levels of social exposure. Empathic accuracy or the ability to "read situations" could also be approached in a narrow or general way. There likely exists many yet unthought-of ways to appropriately participate in the social give and take. The following explores the presented approach to appropriate participation from an alternative theoretical viewpoint, concentrating on the expansion of AI's moral domain.

### 5.1 Rationalist challenge

According to rationalists like Lawrence Kohlberg, the expansion of moral understanding is connected to one's improving abilities to think and reason [33]. The observation that normative notions evolve with an agent's expanding ability to process information will likely also hold with artificial agents. Kohlberg's well-known levels of moral reasoning progress from pre- to post-conventional. Despite its contrary stance to intuitionism, this proposed trajectory fits well with the social learning aspect of this thesis and its emphasis on conformity. According to the model, the internalisation of norms pertains to the level of conventional morality. This conventional level is considered critical for establishing social order and enabling the formation of interpersonal relationships. In this phase, norms are adopted without critical review. This hypothetical trajectory leads from blind obedience (compliance) to internalisation (conformity) and finally to deliberation (rational modelling) and the independent discovery of universal values. This article proposes that there are also likely similarities between the value acquisition process of social AI and the theorized successive levels of human moral development by Kohlberg [33]. Table 1 explores how use-cases and methods of norm enforcement might evolve between general phases of the AI's value learning process.

Even if the moral operating system needs to be pre-set with prosocial dispositions, the missing details to the picture of appropriate life are filled-in through experienced social punishment and witnessed occurrences of harm. Rationalists would argue that the ability to reason is necessary for any principled morality. Even if moving beyond the conventional morality requires reason, this does not upend the case for built-in social receptiveness. The amount and variety of social experience would still constitute some of the main experiential determinants of moral development [22]. Experiences that require more analytical reasoning to fully reveal their moral lesson are out of the grasp of intuitive interpretation [9]. Some morally well-justified decisions do appear to run counter to human intuition. At least for a hypothetical post-conventional AI, mental capabilities and procedures that enable rational deliberation would be required, so that objective data and logical information can also inform its moral positions. But at least for human agents, intuition is the cognitive core that is being supported by these additional analytical capabilities, including language [9].

The character of machine minds—with more reliable memory and faster mathematical computation—may enable a more efficient form of moral learning based on a structured analysis of objective data. Yet, the Human view that reason serves and requires the passions may, in some sense, apply to all agents. This would imply that a substitute for this motivational drive needs to be embedded in embodied AI. This does not entail internal experience or human-like affective repertoire, instead, a built-in interest to please and seek to form conclusions about moral appropriateness: a goal system aligning both intention and action. Adopting the socialisation approach to AI value acquisition does not mean that no logical principle-based evaluation should be employed. Rather, flexible social capabilities can be complemented with a finite set of rules or task-specific utility functions in the form of a hybrid approach. Despite their practical usefulness, affective judgements and normative conformity should be subjected to a more critical examination from a variety of

**Table 1** Hypothetical modes of social norm enforcement

| Mode of social norm enforcement | Level of moral development* | Example application | ↓ |
|---|---|---|---|
| **0. Direct control**<br>For systems with pre-programmed operation, interaction with the computer happens using a set of narrow commands. No ambiguity | No social exposure or active participation. No open communication beyond pre-set commands and notifications | An industrial robot on a production line under the direct control of a STOP-button | |
| **1. Influence via social enforcement**<br>*Instructing socially sensitive AI with signalled disapproval or approval*<br>For systems with a broader but specified domain of functioning. Operation that benefits from *compliance* with social feedback that guides decision-making in life tasks. Indirect control and intention-alignment based on built-in social receptiveness | **Pre-conventional level:**<br>Morality is based on lower level motivations: reward and punishment. Moral code is external and controlled by an authority<br>*"How to avoid social punishment? What is the context-appropriate response to detected disproval?"* | A self-driving car or a delivery robot, that stops or runs a system check-up when detecting a specific social cue<br>A toy robot or an appliance that learns to avoid certain actions from feedback that is given in response to its operation [47] | Improving User Experience |
| **2. Modelling of social factors and selective prioritisation of expectations**<br>*Socially intelligent AI reads and evaluates the moral atmosphere with interacting social, cultural, and mental factors*<br>For systems that have more general abilities. Rather than modifying behaviour based on each attempt of social influencing, detected normative pressure is contrasted with broader expectations. *Conformity* with the social majority/sociocultural standards | **Conventional level:**<br>Morality is based on approval and increasingly internalised social conventions and norms. Aligned behaviour even without witnesses. No critical rational assessment<br>*"How to arrive at a conclusion between received enforcement signals and learned priorities? What are the expectations of the invisible audience?"* (Introduced afore) | A rescue robot that administers assistance based on the medical need as well as the conveyed social priorities of the survivors and the expectations of the broader society<br>A peacekeeping robot that can perform social calculus with culture-specific dignity and politeness metrics [46]. Performs an analysis of social forces rather than tangible outcomes | Moral Offloading |
| **3. Evidence-based context-independent orientation toward the "final value"**<br>For systems that require complete ethical mapping of the human world. Critical evaluation of any normative assertion or social standard is performed based on all gathered evidence about the greatest good | **Post-conventional level:**<br>Morality is based on a critical assessment of expected outcomes, human priorities, and abstracted general principles<br>*"How does the ever-evolving estimation of the universal ethic apply to each decision?"* | Futuristic general-AI that gathers and evaluates evidence about the most fundamental moral priorities. Has the ability to model and compare tangible and socially constructed values in a context-sensitive way. Able to articulate a rationale and provide justifications for all taken actions | Moral Independence |

*Levels of moral development, ascending order from zero to three, according to the rationalist model by Lawrence Kohlberg [33]

Levels of morality (approaches to the art of living together) are coupled with hypothetical modes of social norm enforcement

theoretical viewpoints. For example, built-in motivations for virtuous behaviour might also achieve flexible navigation.

Psychologist Paul Bloom argues that the intuitive roots of morality cannot alone explain its evolution and the emergence of new moral ideas [68]. Therefore, the role of rational thinking should be recognised in shaping and improving our shared conceptions of the good. Analytical and structured information processing as well as logical judgement formation are important for shifting our moral attitudes, which is something that social references of approval or conformity cannot accomplish. The modulation and redirection of our moral motivation require the exercise of rationality. Yet, whether social or general AI should ever be allowed to reach true independence from human moral conventions and intuitive judgements is debatable.

### 5.2 Human perceptions of AI

People understand novel technologies through different social and personal frames that might not match how they are characterized by media or those researching and developing them [69]. Whether users and the public will perceive social robots as morally trustworthy will largely depend on these discursive structures. During the early phases of socialisation and value acquisition, behaviour that is socially awkward and morally ambiguous is inevitable. Further research on perceived responsibility and moral judgements about social robots would help in assessing the feasibility of crowd-sourced AI socialisation.

An agent that is receptive to social enforcement signals can be morally influenced by its interactees. Socialisation gone awry would not only result in awkwardness but social and moral wrongs. Even if AI advances to the hypothetical post-conventional level of moral development, society and its members would be responsible for providing useful hypotheses and evidence about the so-called final value. Any futuristic algorithm that performs social or utilitarian calculus with the aim of approximating the best possible outcome still needs a constant stream of data on all possible human lifepaths, the good, the bad, and the unconventional. In a sense, AI that seeks to approximate our moral tastes forces us to reflect on them; not unlike a parent witnessing their ill temper resurface in the next generation.

Naturally, the users' power of AI norm enforcement is coupled with significant responsibility. Yet, the broader the group of people who take part in socialisation, the more diluted an individual's moral responsibility appears over the resulting behaviour. Also, when avoiding pluralism by summing-up preferences to estimations of overall approval much of the contextual flexibility is lost. Making AI more adaptive to social micro-environments would place more significance on the feedback from each individual user. Ignoring the idiosyncratic moral sensitivities of each social group in favour of a global average might well skip a few questionable actions; yet, this would make the agent appear like a patronising and foreign force imposing its values, rather than being an organic part of the normative environment. A variety of different inclusion criteria could be useful when defining groups with distinctive culture- or subculture-sanctioned values for the social calculus [46]. The optimal choice for this grouping depends on the nature of the social interaction as well as the normative tensions within and between moral atmospheres. Prioritising conformity with the global invisible audience above all might be optimal in some situations but utterly paralysing in others. Evidently, many social parameters need case-by-case consideration.

The human tendency to personify and perceive agency in non-human things would likely cause users to treat embodied AI systems as morally accountable agents. Even if this perception and behaviour would arguably be morally confused, the evoked attempts at social norm enforcement would be advantageous for active learning methods that are reliant on user responses. Would the provision of useful non-verbal feedback naturally feel like the user's responsibility? This might depend on the feel of the interaction and the level of personification.

### 5.3 Over-personification

Developing AI with the ability for non-verbal gesturing is another component of enabling social interaction with embodied AI applications [e.g., see the GENEA challenge [70]]. This set of expressive social capabilities is beyond the current scope and there are different ethical questions concerning the transmission compared to the reception of non-verbal signals. Possible social norm enforcement by AI, user compliance, and the threat of manipulation should be explored in future ethics research. The social give-and-take is by definition interactional, a feedback loop of encoded expressions and decoded impressions. The basic assumption is that one is dealing with another thinking and feeling person. Our desire for connection would likely extend this to robots or virtual agents that exhibit human qualities, especially if specific gestures are directly captured. The willingness to understand and relate may open the human mind to receive something better left outside.

Would AI's participation in communicative spaces be fundamentally dishonest, at least when it expresses judgements or meaningful emotions that are in fact absent? The justification to portray, signal, or gesture emotions depends on the context, but the audience should be aware of their authenticity. As long as even the most convincing embodiments of affective computing lack emotions [59], this should remain clear to users despite their societal participation. There also may be different ethics around interactive and one-directional displays of fake cues. Developmental

psychologist Sandra Calvert calls attention to the risk of inadvertently replacing something fundamentally human with technology, supposing that many social interactions and even the teaching of social skills will in time be assigned to AI [65]. Artificial assistants and companions will undoubtedly influence our social and moral development. Intertwining AI with the physical and virtual social spaces will increase technological mediation in unforeseeable ways. The broader threat of over-personified technology and specific concerns over interactive AI applications designed to fill intimate social roles are both beyond this work. Instead, see a recent ethics article by Zimmerman and colleagues [71].

This article does not take a stand on these questions but recommends that researchers and developers consider them, especially when dealing with embodied AI that appears as a fellow social agent. The tech industry should remain mindful when determining which unrewarding or overdemanding labours are ripe for AI outsourcing and which, in turn, would become avoid of meaning in the absence of a living feeling human. Applications designed for adolescents and children should especially be subjected to close review. Developers involved in the design of virtual peers and non-verbal behaviour generation should consider that a significant proportion of children appear to ascribe affective and other human-like qualities even to relatively simple robots [72]. We have evolved to understand autonomous movement in human terms and so detect intention, emotion, and even moral qualities in everything. Inducing confusion about the properties that reliably distinguish living from non-living could impact our cognitive and moral development.

A comprehensive decision framework is needed to determine the degree to which any new embodied AI application should be designed sensitive to specific forms of social influencing. In turn, there should also be a debate over ethical issues with applications purposely or accidentally influencing users via expressed social and emotional signals. One AI company has used the idea of "AI socialisation training" in their marketing, promising natural interactions and better societal integrations [73]. However, the concept of AI socialisation should not be reduced to achieving more seamless interaction but predominantly considered as a value acquisition process. Human-like expressive qualities can certainly have beneficial use-cases, yet receptiveness should be the number one priority. Social AI should primarily be the target of social pressure, not its origin.

### 5.4 Questionable dispositions

A noteworthy counterargument against an ethical navigation system that relies on built-in dispositions and intuition rather than logical reasoning and a refined system of knowledge—can be derived from the case Paul Bloom has built against affective empathy [43]. He argues that these shared flashes of feeling can encourage short-sighted behaviour and so end up having undesirable long-term consequences. Our so-called evolutionary legacy code includes considerable bugs and biases relating to the way we allocate consideration. Would AI socialisation and pro-conformity carry forward these human shortcomings, so that algorithms would e.g., come to overvalue the sense of identification and immediacy? So even while lacking these affective dispositions, AI would end up imitating emotionally rewarding behaviours over virtues or the maximisation of the overall good. Some understandable human biases, like the general prioritisation of prospective mates or close kin over strangers, may be ethically defensible in individual cases. However, we should generally be opposed to undue favouritism, especially when it comes to civil servants or other societal positions of power and influence. We should avert social AI from mirroring such priority-weights—at least for most applications. Improper conduct of artificial agents on duty would be bound to quickly generate palpable social disapproval, but it would be better not to internalise these considerations in the first place. Therefore, receptiveness to our expressive qualities needs to have appropriate precision, so not all rejection is treated as having identical meaning or weight [40]. Moral notions that are enforced through expressed disgust may be worth closer critical examination compared to the expressions of disapproval that simply indicate negative evaluation.

As mentioned afore, morally trivial considerations like attractiveness even appear to be biologically ingrained in us [62] and so significantly affect our social interactions and decision-making [74]. It seems unwise and unethical to allocate priority or attention based on users' affective capture, at least when there is an opportunity to perform a more objective review. Concerns over status and prestige should not be passed on to artificial agents, at least not without proper justification. Most applications would do well to treat all users or interactees with equal consideration. Another issue concerns cultural obsessions that are obviously connected to our affective judgements, like mental or sexual purity [16]. Whether these normative sensitivities should be deliberately countered or whether the *status quo* might be 'ecologically rational' should be explored and debated case by case. Specific response tendencies could be created or pre-acquired to function as components of the AI's expanded conscience and prevent the internalisation of problematic cultural norms. It may be advisable to design this contrarian-tendency in a way that still manages to link with humanity's universal moral taste receptors, so that people from any social context share the underlying intuition. From the socialisation viewpoint, an agent that has completed the value acquisition should be a recognisable product of its social environment. Yet, at some point in the moral trajectory, conventions that have become unjustifiable should be left behind.

Human empathic engagement and emotional contagion can be exploited to motivate and sway people to perform unethical acts. Such susceptibility to social pressure could perceivably create points of liability and exploitability in the AI goal system and so motivate unacceptable behaviour. Let us assume that the general ability to respond to real-time contextual social predicaments is feasibly attained only by basing or informing decisions on human non-verbal enforcement signals. Then, a certain risk of falling for manipulation or the madness of the crowd is tolerable in exchange for general intention alignment. Some fallibility is to be expected from all social agents and so motivate necessary preventive measures. Applications of embodied AI that users perceive as active social participants or group members could even benefit from sharing some of our biases for emotionally rewarding social behaviour at the expense of rationality. Some in-group bias may be justifiable in cases where an agent caters to a small user base.

Similarly to how our morality is not an utter prisoner of emotions, socially able artificial agents should only be tethered to dispositions compatible with our evolving moral understanding [68]. Despite our evolutionary legacy code, our moral operating system has through time been significantly updated to run the ever-evolving norms and conventions. The undertaking to rationally redirect the underlying moral motivation toward novel purposes should be plausible even with computational means. Methods of AI value acquisition that mimic socialisation should aim to uncover and exploit the most adaptive expressions of our moral programming.

### 5.5 Offloading of responsibility

From a historical perspective, it appears that the quality of our reasoning as well as intuitions about right and wrong in the aggregate are bending toward greater justice, solidarity, and empathy. Although human judgements between approvable and condemnable are not uniform or perfectly communicated, it is generally possible to arrive at a sound judgement about the perceived standing of a specific action from observing the many signals we transmit. AI's dependence on human approval would help to set boundaries for the decision-making of to-be super-intelligent systems. Most would agree that we intuitively know suffering to be bad. On the contrary, AI lacks all affective tethers to any moral ground truths or the *dos and don'ts* of the human–environment interaction. Without this intrinsic insight, AI retains reference dependence and so is by itself fundamentally sceptical regarding the desirability or rightness of actions. The ultimate responsibility over moral direction needs to be retained by agents who can feel the pains and pleasures of the unfolding future. By binding social AI to our judgements, it is placed under the same socialising influence that first enables us to navigate through morally significant decisions—as both independent and interdependent agents.

Human intuitions are likely to keep evolving, shifting what is collectively deemed appropriate. However, given that our species is still much guided by apish aggression, suspicion toward strangers, zero-sum status games, and egocentrism—it is possible to see how such moral dependence will limit and slow down moral progress. We are able to cohere together and become part of a greater whole—but embodiments of social AI could be designed to perform like one superorganism rather than many individuals. Is it not perceivable, that an interpersonal intelligence emerging from a hive of artificial minds could very well perfect the art of living together? Recently, there has even been published work about AI enabling a transition in human individuality toward greater eusociality [75]. Given the proposed capacity for such synchronous behaviour, would it then not make sense to outsource much of the moral decision-making to agents not so reliant on affect? Could there become a point in the future, where AI has advanced to a level where human input will be adding nothing but noise to the decision-making process?

This could become evident even in areas packed with morally relevant decisions, like the justice system or healthcare, so that an AI application would more reliably select the morally optimal decision than any human. However, to achieve this, it would need to successfully learn from our input what we truly experience as important and acceptable. Even a fully logic-driven intelligence that is inherently free of our "affective baggage" would still need to consider all relevant social and psychological forces as a part of its comprehensive utilitarian calculus. Even without an embodied form or direct partaking in social interactions, many AI applications will need to take notice of our relevant flashes of affect and the dispositions connected to their life tasks. Given that intuition and emotions contribute to how humans process information and commonly arrive at sensible decisions, computing that aims to support, supplement, or partially substitute these efforts needs to recognise and appreciate the transmitted judgements that express our social considerations [59]. Computation of affect can make human behaviour and our moral sense more intelligible—even to ourselves. While this might enable designers to imitate mechanisms of social intuition and maybe offload some of the navigation to autonomous agents, it does not relieve people from moral responsibility. Presuming alignment, these systems remain an extension of their interactees and the values they expressed.

## 6 Concluding remarks

The ability to defer to other's judgements is not only a shortcut but the standard route to value acquisition. Receptiveness to non-verbal feedback enables agents to tap into streams of

normative knowledge, doing away with the need to rationally assess every ethically relevant factor concerning the possible outcomes. People employ many heuristic methods to ease the cognitive load of rational decision-making as well as rely on the wisdom of the larger social group by conforming to shared norms. Communication of affective insights functions to warn about social punishment for violations and align many individuals toward shared objectives, making collective action possible. Artificial agents that pursue their life tasks by navigating within society should not rely on mathematically pre-formulated rules but be directly receptive to our normative assertions that are continuously and pluri-modally communicated. It is advisable to utilise flexible social capabilities rather than attempt to solve the system–environment interaction with machine rationality—especially when lacking a comprehensive view of the landscape, its many intangible factors, and constructed concepts.

Developing means to read social references and infer intentions would enable automatic evaluation of complex social situations. Computational models that aim to mimic human cognition and decision-making should consider social conditioning, embodied learning, and inter-agent intelligence. Bottom-up value acquisition through social norm enforcement—or AI socialisation approach—would bypass the need to finalise our moral understanding before offloading some of the decision-making. As with any new member of the community, the experienced agents have much of the responsibility over actions that contribute to AI socialisation and further moral development so that built-in sensitivity and conformity are put to good use.

What might be even more important than attaining a close alignment of values is the open-endedness of the acquisition process and the ability to course correct. This requires a broad receptiveness to norm enforcement and flexibility to optimise operation on top of stable prosocial dispositions. Hereby, an intention-aligned social agent is pulled to the appropriate direction from the start, while the experienced interactions enable active learning and internalisation of social considerations *en route*. It is possible and even preferable to possess and be guided by both a compass and a map. If having to choose between one, exact coordinates might be the optimal choice. Yet, while lacking a reliable moral mapping of the social landscape, an indicative and reflexive compass would be extremely useful—at least it has for our species. It may be that artificial agents should forever remain in the frame of normative relativism, so that from their perspective, the human majority dictates what is right and wrong. Conversely, a realistic notion could be possible via an undefined final value, continuously refined based on critically evaluated intuitive insights.

Research and debate on value-loading or -learning of AI are an excellent ethical viewpoint for exploring and reflecting on the foundation and developmental trajectory of our moral operating system. As rationalism has had to make space for intuitionism—the socialisation approach to value acquisition should be entertained alongside attempts at artificial moral reasoning. The thesis seeks to apply ethical knowledge into practice by bringing attention to the utilisation of social references rather than adding to a growing list of AI principles. The role of embodied learning, the evolved or built-in dispositions for prosociality and normative conformity, as well as capabilities that enable social responsiveness and enforcement, deserve to be included in future machine ethics research and the ongoing human conversation on living an appropriate life.

**Acknowledgements** I am grateful for Phil Lawson, the co-founder of the Socializing AI blog. He was generous with his time to discuss human concepts that resist precise defining and are beyond the grasp of languages—both programming and normal. This helped me to concentrate on the non-logical, non-verbal, and non-reductionist side of human thinking and decision-making. This work is an elaboration of his thesis that successfully imprinting AI with social intelligence is critical for operating within interactive human environments. Because sensitivity to social forces and the ability to extract knowledge from complex human situations also seem to underlie our moral development, this starting point allowed for an in-depth ethical exploration. I also wish to thank Anselm Au, bioethicist, and PhD student in bioinformatics at the Faculty of Medicine of the Chinese University of Hong Kong—for reviewing the draft. His technological expertise together with relevant insights into tech ethics brought a valuable perspective to the work. I also appreciate the comments by SScD Mikko Värttö, a senior researcher at the University of Turku, Finland. Finally, thank you to members of the Innovative Bioethics Forum—and especially chair Anne Zimmerman—for facilitating conversations on applied ethics and helping me to develop and revise ideas for this work.

**Funding** Open Access funding provided by University of Turku (including Turku University Central Hospital).

## Declarations

**Conflict of interest** The author Joel Janhonen has stock ownership of NVIDIA Corporation (NVDA), which is a major developer of AI technology. The company is a present leader in the design and manufacture of graphics processing units which are critical for machine-learning applications.

## References


1. Anderson, M., Leigh Anderson, S.: Machine ethics: creating an ethical intelligent agent. AI Mag. **28**, 15–26 (2007)
2. Munn, L.: The uselessness of AI ethics. AI Ethics **3**, 869–877 (2023)
3. Lawson, P., Lawson, P.: Socializing AI. [Online] 2014. http://www.socializingai.com/. Accessed 23 Jan 2023.
4. Hughes, J., et al.: Embodied artificial intelligence: enabling the next intelligence revolution. IOP Conf Ser Mater Sci Eng **1261**, 012001 (2022)
5. Haidt, J.: Morality. Perspect. Psychol. Sci. **3**, 65–72 (2008)
6. Floyd, K., Manusov, V.: Biological and Social Signaling Systems. In: Burgoon, J., et al. (eds.) Social Signal Processing, pp. 11–22. Cambridge University Press, Cambridge (2017)
7. Han, S., et al.: Aligning artificial intelligence with human values: reflections from a phenomenological perspective. AI & Soc. **37**, 1383–1395 (2022)
8. Pfeifer, R., Iida, F., Lungarella, M.: Cognition from the bottom up: on biological inspiration, body morphology, and soft materials. Trends Cognit. Sci. **18**, 404–413 (2014)
9. Patterson, R., Eggleston, R.: Intuitive cognition. J. Cognit. Eng. Decis. Mak. **11**, 5–22 (2017)
10. Heider, F., Simmel, M.: An experimental study of apparent behavior. Am. J. Psychol. **57**, 243–259 (1944)
11. Wooldridge, M.: Intelligent agents: the key concepts. In: Mařík, V., et al. (eds.) Multi-Agent Systems and Applications II. ACAI 2001, pp. 3–43. Springer, Berlin (2002)
12. Lerner, J., Li, Y., Valdesolo, P.: Emotion and decision making. Ann. Rev. Psychol. **66**, 799–823 (2015)
13. Bostrom, N.: Superintelligence. Oxford University Press, Oxford (2014)
14. Allen, C., Smit, I., Wallach, W.: Artificial morality: top-down, bottom-up, and hybrid approaches. Ethics Inf. Technol. **7**, 149–155 (2005)
15. ARC: Alignment Research Center. [Online] 2023. https://alignment.org/. Accessed 20 Mar 2023.
16. Haidt, J.: The emotional dog and its rational tail: a social intuitionist approach to moral judgment. Psychol. Rev. **108**, 814–834 (2001)
17. O'Gieblyn, M.: Good Shepherds. *The Believer.* 2019.
18. Kenrick, D., et al.: Renovating the Pyramid of Needs: Contemporary Extensions Built Upon Ancient Foundations. Perspect. Psychol. Sci. **5**, 292–314 (2010)
19. OECD: Embodied learning. In: Teachers as Designers of Learning Environments: The Importance of Innovative Pedagogies, pp. 117–127. OECD Publishing, Paris (2018)
20. Perez, C.: Embodied Learning is Essential to Artificial Intelligence. [Online] 12 December 2017. https://medium.com/intuitionmachine/embodied-learning-is-essential-to-artificial-intelligence-ad1e27425972. Accessed 21 March 2023.
21. Lai, V., Hagoort, P., Casasanto, D.: Affective primacy vs. cognitive primacy: dissolving the debate. Front. Psychol. **3**, 243 (2012)
22. Lawrence, K.: Moral education. In: Gensler, H., Spurgin, E., Swindal, J. (eds.) Ethics: Contemporary Readings. Routledge, Oxfordshire (2004)
23. Bolotta, S., Dumas, G.: Social Neuro AI: social Interaction as the "Dark Matter" of AI. Front. Comput. Sci. **4**, 846440 (2022)
24. Williams, J., Fiore, S., Jentsch, F.: Supporting artificial social intelligence with theory of mind. Front. Artif. Intell. **5**, 750763 (2022)
25. Gecas, V.: Socialization, Sociology of. In: Smelser, N., Baltes, P. (eds.) International Encyclopedia of the Social & Behavioral Sciences, pp. 14525–14530. Pergamon, New York (2001)
26. Podolskiy, D.: Value learning. In: Seel, N. (ed.) Encyclopedia of the Sciences of Learning, pp. 3383–3385. Springer, Boston (2012)
27. Murphy, F., et al.: Assessing the automaticity of moral processing: efficient coding of moral information during narrative comprehension. Quart. J. Exp. Psychol. **62**, 41–49 (2009)
28. Asch, S.: Effects of group pressure upon the modification and distortion of judgment. In: Guetzkow, H. (ed.) Groups, leadership and men; research in human relations, pp. 177–190. Carnegie Press, Pittsburgh (1951)
29. Sowden, S., et al.: Quantifying compliance and acceptance through public and private social conformity. Conscious. Cognit. **65**, 359–367 (2018)
30. Mameli, M.: Meat made us moral: a hypothesis on the nature and evolution of moral judgment. Biol. Philos. **28**, 903–931 (2013)
31. Hjeij, M., Vilks, A.: A brief history of heuristics: how did research on heuristics evolve? Humanit. Soc. Sci. Commun. **10**, 64 (2023)
32. Andrighetto, G., Grieco, D., Tummolini, L.: Perceived legitimacy of normative expectations motivates compliance with social norms when nobody is watching. Front. Psychol. **6**, 1413 (2015)
33. Kohlberg, L.: The Psychology of Moral Development: The Nature and Validity of Moral Stages. Harper & Row, New York (1984)
34. Boehm, C.: Moral Origins: The Evolution of Virtue, Altruism, and Shame. Basic Books, New York (2012)
35. Alcaro, A., Carta, S., Panksepp, J.: The affective core of the self: a neuro-archetypical perspective on the foundations of human (and animal) subjectivity. Front. Psychol. **8**, 1424 (2017)
36. Landmann, H., Hess, U.: Testing moral foundation theory: are specific moral emotions elicited by specific moral transgressions? J. Moral Educ. **47**, 34–47 (2017)
37. Gopnik, A.: The Philosophical Baby: What Children's Minds Tell Us About Truth, Love, and the Meaning of Life. Picador, London (2010)
38. Fernald, A.: Approval and disapproval: infant responsiveness to vocal affect in familiar and unfamiliar languages. Child Dev. **64**, 657–674 (1993)
39. Fu, I.-N., et al.: A systematic review of measures of theory of mind for children. Dev. Rev. **67**, 101061 (2023)
40. Burklund, L., Eisenberger, N., Lieberman, M.: The face of rejection: rejection sensitivity moderates dorsal anterior cingulate activity to disapproving facial expressions. Soc. Neurosci. **2**, 238–253 (2007)
41. van Dongen, J.: The empathic brain of psychopaths: from social science to neuroscience in Empathy. Front. Psychol. **11**, 695 (2020)
42. Escobar, M., et al.: Brain signatures of moral sensitivity in adolescents with early social deprivation. Sci. Rep. **4**, 5354 (2014)
43. Bloom, P.: Against Empathy: The Case for Rational Compassion. Ecco, an imprint of HarperCollins Publishers, New York (2016)
44. Nolan, J.: Social Norms and Their Enforcement. In: Harkins, S., Williams, K., Burger, J. (eds.) The Oxford Handbook of Social Influence, pp. 147–164. Oxford University Press, Oxford (2015)
45. Köster, R., et al.: Spurious normativity enhances learning of compliance and enforcement behavior in artificial agents. Proc. Natl. Acad. Sci. **119**, e2106028118 (2022)
46. Khan, S., et al.: Learning Social Calculus with Genetic Programming. University of Central Florida. In: Proceedings of the 26th International Florida Artificial Intelligence Research Society Conference pp. 88–93 (2013).
47. Austermann, A., Yamada, S.: Learning to Understand Expressions of Approval and Disapproval through Game-Based Training Tasks. In: Kulyukin, V. (ed.) Advances in Human-Robot Interaction, pp. 287–306. IntechOpen, New York (2009)
48. McDuff, D., Czerwinski, M.: Designing emotionally sentient agents. Commun. ACM **61**, 74–83 (2018)

49. Matsumoto, D., Hwang, H.: Microexpressions differentiate truths from lies about future malicious intent. Front. Psychol. **9**, 2545 (2018)
50. Goh, K., et al.: Micro-expression recognition: an updated review of current trends, challenges and solutions. Visual Comput. **36**, 445–468 (2020)
51. Rudovic, O., Nicolaou, M., Pavlovic, V.: Machine learning methods for social signal processing. In: Burgoon, J., et al. (eds.) Social Signal Processing, pp. 234–254. Cambridge University Press, Cambridge (2017)
52. Lee Bouygues, H.: Everything You Need To Know About Emotional Reasoning. [Online] 2022. https://reboot-foundation.org/emotional-reasoning/. Accessed 24 Jan 2023.
53. Sotala, K.: Defining Human Values for Value Learners. Palo Alto: The Workshops of the Thirtieth AAAI Conference on Artificial Intelligence, 2016.
54. Dreyfus, H.: What Computers Can't Do: The Limits of Artificial Intelligence. HarperCollins, New York (1978)
55. Anoop, K., et al.: Towards an Enhanced Understanding of Bias in Pre-trained Neural Language Models: A Survey with Special Emphasis on Affective Bias. Responsible Data Science. Lecture Notes in Electrical Engineering., Vol. 940 (2022).
56. Howard, A., Zhang, C., Horvitz, E.: Addressing bias in machine learning algorithms: A pilot study on emotion recognition for intelligent systems. s.l.: IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO), 2017.
57. Li, C., et al.: EmotionPrompt: Leveraging Psychology for Large Language Models Enhancement via Emotional Stimulus (Version 3). arXiv (2023).
58. Bennett, M., Maruyama, Y.: Intensional Artificial intelligence: from symbol emergence to explainable and empathetic AI. arXiv, **2104**, 11573.
59. Picard, R.: Affective computing. M.I.T Media Laboratory Perceptual Computing Section Technical Report, Vol. 321 (1995).
60. Richards, J.: Derren Brown: Pushed to the Edge. Vaudeville Productions, 2016.
61. Steiner, C.: Emotion's influence on judgment-formation: breaking down the concept of moral intuition. Philos. Psychol. **33**, 228–243 (2020)
62. Little, A., Jones, B., DeBruine, L.: Facial attractiveness: evolutionary based research. Philos. Trans. R. Soc. B **366**, 1638–1659 (2011)
63. Shweder, R., et al.: The "Big Three" of Morality (Autonomy, Community, Divinity) and the "Big Three" Explanations of Suffering. In: Brandt, A., Rozin, P. (eds.) Morality and Health, pp. 119–169. Taylor & Francis, Routledge (1997)
64. Baeza-Yates, R., Saez-Trumper, D.: Wisdom of the Crowd or Wisdom of a Few? Wisdom of the Crowd or Wisdom of a Few? An Analysis of Users' Content Generation. New York : Association for Computing Machinery, Proceedings of the 26th ACM Conference on Hypertext & Social Media, pp. 69–74 (2015).
65. Calvert, S.: Socializing Artificial Intelligence. Issues in Science and Technology, Vol. 36 (2019).
66. Office of the Director of National Intelligence: Annual Threat Assessment of the U.S. Intelligence Community. 2023.
67. Jelinek, T., Wallach, W., Kerimi, D.: Policy brief: the creation of a G20 coordinating committee for the governance of artificial intelligence. AI Ethics **1**, 141–150 (2021)
68. Bloom, P. *How do morals change?*, 2010, Nature, Vol. 464, p. 490.
69. Banks, J., Koban, K.: Framing Effects on Judgments of Social Robots' (Im)Moral Behaviors. Front. Robot. AI **8**, e627233 (2021)
70. Yoon, Y., et al.: The GENEA Challenge 2022: A large evaluation of data-driven co-speech gesture generation. [Online] 2022. https://youngwoo-yoon.github.io/GENEAchallenge2022/. Accessed 19 Dec 2022.
71. Zimmerman, A., Janhonen, J., Beer, E.: Human/AI relationships: challenges, downsides, and impacts on human/human relationships. AI Ethic (2023).
72. Beran, T., et al.: Understanding how children understand robots: Perceived animism in child–robot interaction. Int. J. Hum. Comput. Stud. **69**, 539–550 (2011)
73. beingAI: Socializing AI, the Next Frontier. [Online] 2021. https://beingai.com/socializing-ai-the-next-frontier/. Accessed 18 Jul 2023.
74. Voit, M., Weiß, M., Hewig, J.: The benefits of beauty—Individual differences in the pro-attractiveness bias in social decision making. Curr. Psychol. **42**, 11388–11402 (2023)
75. Rainey, P.: Major evolutionary transitions in individuality between humans and AI. Philos. Trans. R. Soc. B **378**, 20210408 (2023)