Towards a Socio-Legal Robotics: A Theoretical Framework on Norms and Adaptive Technologies

Larsson, Stefan; Liinason, Mia; Tanqueray, Laetitia; Castellano, Ginevra

Published in:
International Journal of Social Robotics

DOI:
10.1007/s12369-023-01042-9

2023

Document Version:
Publisher's PDF, also known as Version of record

International Journal of Social Robotics
https://doi.org/10.1007/s12369-023-01042-9

Towards a Socio-Legal Robotics: A Theoretical Framework on Norms and Adaptive Technologies

Stefan Larsson[1] · **Mia Liinason**[2] · **Laetitia Tanqueray**[1] · **Ginevra Castellano**[3]

Accepted: 3 August 2023

Abstract
While recent progress has been made in several fields of data-intense AI-research, many applications have been shown to be prone to unintendedly reproduce social biases, sexism and stereotyping, including but not exclusive to gender. As more of these design-based, algorithmic or machine learning methodologies, here called *adaptive technologies*, become embedded in robotics, we see a need for a developed understanding of what role social norms play in social robotics, particularly with regards to fairness. To this end, we (i) we propose a framework for a *socio-legal robotics*, primarily drawn from Sociology of Law and Gender Studies. This is then (ii) related to already established notions of acceptability and personalisation in social robotics, here with a particular focus on (iii) the interplay between adaptive technologies and social norms. In theorising this interplay for social robotics, we look not only to current statuses of social robots, but draw from identified AI-methods that can be seen to influence robotics in the near future. This theoretical framework, we argue, can help us point to concerns of relevance for questions of fairness in human–robot interaction.

Keywords Socio-legal robotics · Human-Robot Interaction · Social norms · Gender studies · Mirroring of norms · Adaptive technologies

1 Introduction: The Social in Social Robotics

In this study, we theorise on the implications of social norms for social robotics, informed by gender studies and socio-legal theory. We use recent AI-developments as an opportunity to demonstrate and emphasise the need for this, although our scope includes also non-learning algorithmic systems and robotic design as such.

Social robots are, most simply put, robots that can interact and communicate with humans (cf.[1]). Scholars within the realm of Human–Robot Interaction (HRI) define social robots as having distinctive personality and character traits, as well as perceive and express emotions; those enable social robots to communicate through the use of natural cues—such as gaze and gestures—and 'expected norms' within a given context (cf. [2–4]), and lead to a further need for addressing ethical and legal questions when empowered with AI and autonomy ([5]). Social robots hence take part in a social context, which has prompted studies on what social norms they reflect [6], including with regards to gender [7]. Interestingly, in a scrutiny of gender and robotics, it has been pointed out that actual practices of robotics at worst may serve to "re-entrench existing social stereotypes and hierarchies rather than to contest them" [8, p.360] (cf. [9]). A recent study however found evidence of that breaking gender norms boosts robot credibility regardless of gender or cultural context, and regardless of pretest gender biases [10]. In addition, while recent advancements in AI-research has enabled robots to recognise faces [11], synthesise speech to

Mia Liinason, Laetitia Tanqueray and Ginevra Castellano have equally contributed to this work.

✉ Stefan Larsson
stefan.larsson@lth.lu.se

Mia Liinason
mia.liinason@genus.lu.se

Laetitia Tanqueray
laetitia.tanqueray@lth.lu.se

Ginevra Castellano
ginevra.castellano@it.uu.se

[1] Department of Technology and Society, LTH, Lund University, Lund, Sweden

[2] Department of Gender Studies, Lund University, Lund, Sweden

[3] Department of Information Technology, Uppsala Social Robotics Lab, Uppsala University, Uppsala, Sweden

Published online: 20 October 2023

be more accepted [12], observe and learn from human movement [13] it has also stressed the need for ways to better ensure fairness in HRI. Regarding norms and gender, this need has for example been argued from the basis of how traditional housewife ideals still are reproduced in technological devices for the home [14] and can be seen in calls for *feminist HRI* [15]. That is, as data-dependent AI-systems and large language models further power robots—where the way we speak, write and behave in various contexts becomes training data—there is a continuous need to theoretically revisit the relationship between the social and the robot, also beyond design and "non-learning" systems. Particularly so as social robots make-up a part of a complex, and in some distinct ways problematic, data-driven, sociotechnical world.

Indeed, researchers have found that AI-systems frequently mirror and display existing prejudices and societal injustices [16–19]. There have been numerous awareness-raising scandals where the false universal of 'man' or 'whiteness', taken as coexistent with 'being human' itself, have interplayed with learning technologies. These include commercial facial-recognition systems with much less accuracy for female faces and dark skin [20], gender-discriminating job ads [21], and antifeminist and anti-Semitic sentiments expressed by adaptive conversational agents [22]. Consequently, as long as robotic design at large, including machine learning-based systems that learn from incomplete datasets, depend on biased and unfair social structures, we can expect them to not only reproduce but also amplify inequalities [23–25]. In parallel and despite the best of intentions, there is a risk that definitions of fairness may fail to consider how the social context intermesh with technology in different forms. We use the term *adaptive technologies* (see Fig. 1) to theorise on how robotic design, with or without AI-systems, adapt to or reproduce social norms in their design or via collected data.

In short, we see a need for more interdisciplinary work in social robotics *in relation to* social norms and stereotypes. We use theoretical findings in Sociology of Law (SoL) and Gender Studies to highlight the mirroring of norms in robotic learning and design. While law, as in formal norms, are of relevance for governing technologies, and also adapt in relation to technological development [26, 27], our study mainly focuses informal, social, norms. This focus is often found in socio-legal studies, hence the proposed terminology of *socio-legal robotics*. With regards to adaptiveness, we refer to technologies that purposefully are built to change, transform or develop the relation to social interaction, which social robotics falls into. This is also why we point to *normative mirroring*, a term suggested in the socio-legal literature with regards to the interplay between AI-systems and society [23, 28], as one way to conceptualise this space of scrutiny. Lastly, while we primarily draw from discourses on gender, it should here be seen in its wider intersectional approach. That is, gender is framed in a way that is not solely about the person's sex, but is also impacted by a person's race, age, (dis)ability, ethnicity and socio-economic status to name a few, which all play a role in how a person lives and experiences their everyday life [29–31].

1.1 Purpose and Aims

In this article, we propose a conceptual framework on *socio-legal robotics* through the interplay of SoL, Gender Studies and HRI. As disciplines, SoL and Gender Studies are both critical disciplines to understand the interplay of various norms and power structures in certain settings, often with contributions envisioning a more just society; whilst HRI is a discipline seeking to bring social robots into society. Combining all three disciplines accommodates for emerging concerns in social robotics. That is, to be able to theorise and understand how to deal with the fact that the underlying technologies are, and likely increasingly so, becoming adaptive of the social interplay that includes social norms and informal social structures, e.g. related to gender. Put differently, as robots adapt to and mirror gendered informal structures, a heightened awareness is necessary with regards to understanding the complexities of this interaction.

Under this aim, we seek to:

1. Propose a theoretical basis of socio-legal robotics, primarily drawn from the realms of social sciences that underpins both Sociology of Law and Gender Studies.
2. Relate the theoretical framework to already established notions of acceptability and personalisation found in social robotics as a field, in order to sketch three levels of adaptive technologies: design, datasets and in situ personalisation. That is, if a robot mirrors or adapts to and "learns" social norms, how may we better understand the implications of this mirroring?
3. Contribute to how we can think of fairness as a contextualised and situated practice, in light of robotic mirroring of social norms.

1.2 Developing the Argument: Acceptability and Personalisation

Robots' ability to learn from and adapt to humans is arguably a key aspect in the field of social robotics [32]. With the advent of learning technologies strengthened by data-dependent ML and automation, social robotics is moving beyond pre-programmed rule-based systems, and towards human-in-the-loop ML-based approaches for the generation of what can be called socially adaptive robot behaviours [33]. Correspondingly, official projections predict that in the course of the next two decades, societies will see pervasive use of robotic technology in all contexts of social interaction, public and private [34]. There is development in

techniques for reinforcement learning in social robotics [35], adaptive robotic tutors [36] and various studies on personalisation in anything from learning scenarios to bartending (cf.[37, 38]). In addition, development in language modelling and generative AI like GPT generations—which has been found to be biased towards gender stereotypes [39]—is likely to an impact in HRI as well (cf. a "foundation agent" for robot manipulation, learning from human movements [40]). Despite its importance for social robotics applications, work on context-specific norm learning on robots has been limited [41]. However, research in social robotics has recently begun to take an interest in normative issues from an adaptive point of view. This, for example, concerns 'affective' robotics, arguing for the need for robots to be able to be attentive to moods and attitudes in HRI [42] or how robots may be able to reason about social norms in order to plan appropriate behaviour [43] or how robots may challenge gender norms [44]. This includes a critical perspective on robotics, that has been advocated for as a way to identify conflicting ideas about technologies, particularly with regards to understanding innovations in robotics and their potential social consequences [45].

1.2.1 Robotic Mirroring of Human Traits

Previous research has shown that mirroring human traits in the design of both a robot's appearance and behaviour may be beneficial for acceptability by humans users [7, 46–48]. What we do not know yet is exactly how much or what type of mirroring is necessary for acceptability, as well as how context (i.e. application, culture) dependent this mirroring should—or should not—be, and how it should be accounted for in ML-based approaches to endow robots with socially interactive skills.

More specifically, some research indicate that the more human-like a robot is in appearance, the more likely people are to empathise with it [46]. People's responses change, though, in presence of very human-like robots, such as androids and humanoids. Robotic agents that are very human-like in appearance, but robotic-like in their behaviour, may generate a violation of people's expectations, which evokes feelings of eeriness and unfamiliarity, according to the so called uncanny valley hypothesis [49, 50]. Moreover, researchers have investigated broader issues that relate to the mirroring of human traits in robots and that may be key for the acceptance of robots in society, for example cultural aspects. Šabanović shows this in a critical exploration of how Japan has introduced social robots in society [8]. She demonstrated that researchers are expected to reproduce conservative social values, an "assumed cultural homogeneity", in order for social robots to be accepted by consumers—which is problematic in and of itself [8, p.358].

1.2.2 Gendering Robots

Research on social robots and gender shows that there are still complex issues being tackled around whether and how social robots should be gendered, both in appearance and behaviour, as well as whether robots should mirror gender norms. These are still open questions, and important, especially as we discuss in relation to the mirroring of social norms. The interplay between technological design and gender is a wide field, with several highly problematic biases in both design, designers, and data, pointed to by works from for example Criado Perez [30] and D'Ignazio and Klein [51]. Therefore, on one hand, the lack of a gender-sensitive approach to the design of social robots in certain applications might lead to the phenomenon of gender data gap [30], where there is a lack of representative data for women.

On the other hand, there is a risk that the gendering of robots and thereby mirroring social norms may lead to the reproduction and perpetuation of gender biases. A recent UNESCO report, for example, has demonstrated that the gendering of voice assistants led to production and reproduction of gender stereotypes, especially the notion that women should be submissive, polite and patient [52]. In social robotics, studies have explored the relationship between robot gendering and gender and occupational stereotypes [53, 54], human likeness [55] and perception of robotic non-compliance [56]. Moreover, recent studies investigated how social robots could be designed to go against current digital assistants' gender norms and suggest that feminist robots can play a role in reducing gender biases and harmful stereotyping [44] (see also the mentioned calls for *feminist HRI* [15, 57]). Recent commentary have made compelling arguments on the need to develop the critical scrutiny of the detailed learning techniques in order to better understand discriminatory implications and thereby contribute to a more aware development of robots [58, 59]. Even within the same application scenario, a gender-sensitive approach to the design of social robots pointed to the fact that mirroring some gender aspects might be beneficial (e.g. a robot's appearance), while the mirroring of gendered norms might not [7].

Furthermore, research shows that non-representative datasets used to train ML algorithms for automatic face recognition might lead to disparities in recognition accuracy for under-represented groups [20]. Consequently, new investigations are needed on how gender should be accounted for in datasets used for training robot behaviours for social interactions. We posit that there is a need for more focus on how to handle questions of biases or harmful social norms picked up by robots' interactive skills, and learned via ML-based approaches, e.g. with datasets capturing different types of human interactions. It is currently unclear how the design of more inclusive datasets (e.g. in terms of gender, ethnicity) could lead to more inclusive robotics (cf. [7]), and how

this, instead, might lead to social robots reproducing harmful human biases.

1.2.3 Adaptive Technologies and Norms

There is an increased awareness in the wider community of AI-research around the need to address ethical challenges and questions of norms. For example, an IEEE report on ethical issues linked to autonomous and intelligent systems acknowledges the possibility of "norm conflicts", where for example tension "may sometimes arise between a community's social and legal norms and the normative considerations of designers or manufacturers" [60, p.175]. This calls for a deeper understanding of the interplay between robots, or "adaptive technologies", and social norms, particularly in relation to questions of fairness. In Sect. 3 below, we propose a model for how this interplay can be understood, and in Sect. 4 we conclude by promoting distinct areas of focus in the path ahead.

In sum, it is reasonable to assume that a coming strand of research in social robotics will have to be rooted in an interdisciplinary theorising on what type of norms socially adaptive robots should and should not reproduce, and how this interaction can be better studied and understood from an applied, everyday perspective in relation to social structures and gender. In the following section, we primarily look to SoL and Gender Studies to contribute to a theoretical frame of socio-legal robotics that can guide the understanding of *the social* in social robotics.

2 Theoretical Development: Interdisciplinary Contributions

In this section, we draw from Sociology of Law (SoL) and Gender Studies to theorise on how social norms and stereotypes inform and are mirrored in robotic design, how design can be normative, and what importance situatedness and contextuality have for addressing questions of fairness in robotics.

2.1 Sociology of Law as a Study of Norms

SoL, as a discipline, theorises and empirically studies the relationship between law and society, including non-formal aspects of social control [61]. For many years, SoL scholars have studied legitimacy in terms of social norms (often based on the notion of "social facts" [62], as a set of informal expectations that can be compared or contrasted to law [63–66], and is of relevance to social control [67]. Early accounts point to *living law* [68] or *law in action* [69] to explain what regulates the social life of communities and societies.

Drawing from SoL, we use a definition of norms that are (i) shared (and thereby social), (ii) expectations on behaviour by (iii) groups. Our definition of social norms take on a problematising approach in that we stress what Bicchieri et al. [70] call "scripted" or group behaviour. We focus this particular aspect of social norms which relates to conformity to group norms in a stereotyping way, which has been studied in HRI in elderly care in the sense that stereotypes "set normative expectations about how a good group member should behave" [71, p. 2]. That is, we treat stereotypes as a way for how a particular set of social norms can be expressed. They can both be useful for design—as with the acceptability described above—but also be harmful and unjust. Social norms have been studied in relation to group dynamics and how, in the words of Bicchieri et al., "social identity is built around group characteristics and behavioral standards, and hence any perceived lack of conformity to group norms is seen as a threat to the legitimacy of the group" [70, p. 9]. This is pointing to a social mechanism that can include problematic aspects of how social control may be both stereotyping, harmful, or in other ways, toxic.

We link these "scripted" group behaviours to stereotypes, that reciprocates also in technology development, such as in robotic design [9], in the "smart wives" of the home [14], or in how the sampling of data for large datasets in some instances has been shown to contain misogyny and malignant stereotypes [72]. Traditional gendered stereotypes see traits such as ambition, power and competitiveness as inherent in men, and traits such as nurturing, empathy and concern for others as characteristics of women [73]. These stereotypes can be reproduced also in robotic design. Consequently, and as concluded in a study on recommendations systems, that gender bias issues in AI recommendations cannot fully be addressed without addressing the gender biases in humans [74].

This means—in light of conversational agents adapting to anti-feminist sentiment, being designed to passively accept sexual harassment [52], or employment of gender-biased hiring applications [75]—that we advocate for an increased awareness of the importance of *which social norms* are learnt, also for social robotics [59]. Thus, the challenges with AI-supported robotics, for example, are more nuanced than to be about any simpler form of "alignment" to "human values" (for an extensive account, see [76]). Exactly what values, or norms, that gets "aligned" are far from a consensual, democratic or neutral process. In terms of design, this could mean that whatever perceptions of gender or bodies that the designers have may poorly influence design for women [30] or persons with disabilities [77], for example.

2.1.1 The Robotic Mirroring of Social Norms

Hence, these insights on the mirroring of norms in adaptive technologies lead to a normative question of what social norms a social robot *ought* to or *ought not* to reproduce. In approaching this normative question, we stress the need for an awareness of that such a robotic mirroring of social norms is an important question in the first place. This approach on normativity indicates a scope that goes beyond formal law, as it poses issues which affect society and communities generally in ways which formal law might not account for [78]. This sensitivity for social norms includes many mundane everyday situations that, although clearly structured and guided by social cues and conventions, are not necessarily primarily governed by formal law. This could concern how we converse or communicate in various contexts, including social media, behave at a dinner table, or how different norms on gender affect anything from family structures, partnerships and professional expectations. The legislature admittedly has difficulty regulating human judgment in different life situations, albeit there may of course be legal frames or boundaries surrounding these contexts. But many social norms regarding gender, family, sexuality and relationships, for example, are informal and "unspoken". They are exercised through bodily acts and speech [65] in everyday public and private situations, at work, in school and in the family. The main argument here lies in the sometimes problematic relationship to robotic design and how adaptive technologies make use of data from these contexts.

These social norms may not only be seen as guiding communities in a multitude of non-formal normative issues, but also—of particular relevance here—perpetuate biases and unfair social structures. From a ML-perspective, this can be seen in what is expressed and captured in what images are included in the collection of facial features [20], what data that was used for prediction tools in human resources [79], or the human tagging of images in image-databases [80]. So, social norms expressed in texts in books, emails or on websites, but also organisational structures, gendered labour markets and purchase patterns, are used to train algorithmic models to detect, translate and predict. This is discussed by Larsson [23] with regards to "data-dependent AI that learns from real world examples derived from human activities may be understood as a mirror of social structures, leading to questions of accountability for those devising the mirror, its reproducing as well as amplifying abilities" [23, p. 589]. It is this type of mirroring we advocate needs further scrutiny in social robotics, and develop a model for below.

Pointing back to questions of acceptability, and to add a layer of complexity, there may indeed be useful and functional aspects of "personalised" robots, that can adapt to ways of talking or behaving in order to be accepted [38]. There are likely many non-problematic ways that personalised robotics can be used for developing HRI. However, one can also picture that for some communities or contexts it is stereotypical or even misogynistic expressions that could contribute to human acceptability—in that particular group—for robots, if they thereby mirror norms present in that community or context. Similar issues have been analysed from an ethical perspective in the case of gender-stereotyping in robotic eldercare [71]. This points to the normative complexity of meaning-making, acceptability and robotic mirroring of social norms. A first step, from a critical point of view, is however to acknowledge that this adaptivity may at worst pick up "a number of structural biases and imbalances that societies struggle with in general, such as inequality, unfairness, discrimination and racism" [23, pp.589–590].

In practice, for the adaptive aspects of social robots, it means that social robots may directly reflect society and its various contexts, some of which harmful, discriminatory or violent. Tanqueray et al. [7], for example, have demonstrated that socially assistive robots in the context of perinatal depression may mirror unwanted practices for the screening of peripartum depression, and perpetuate the narrative of the more powerful institutions at play. Furthermore, those who help develop such a technology may overlook gendered power relations and power structures [15, 18, 51]. Correspondingly, Tanqueray et al [7] show that the bridging of SoL and HRI is needed to critically develop social robots in a given context.

2.1.2 Code as Law, and the Materiality of Robotic Norms

Beyond law, norms can also—in a sense—be coded or designed into material objects. In debating how the early Internet met and related to normative structures, the legal scholar Lawrence Lessig argued for code—in terms of "cyberspace"—becoming "law" [81, 82]. For Lessig, this was a way to stress that another type of often overlooked regulator, that is, an entity that actually controls behaviour, in addition to formal legislation. Next to formal law, he also included markets, social norms and architecture as a governance structure (cf. [65, pp.589–590]). This notion of "code as law" has influenced much thought on how governance is played out for primarily digital environments, for example on digital platforms [83], digitally mediated property [84], but also governance of AI, for example in Japan [85], as well as in robotics [86]. This technologically designed side of norms may not be explicitly *intended* to be normative—not in the sense that formal law explicitly is intended to be normative—but may, just as well, be. Consequently, in the critical AI literature, as shown above, there is much critique found in *whom* is developing [17, 30, 51, 52] (see also the call for a feminist HRI [15, 57]) in the sense that this privilege also affects design normatively, often with blind-spots for those groups that are not heard or part of the development [6,

51]. This is also why we include *design* in the three levels of adaptive technologies below, shining a light also on risks of non-diverse sampling in HRI research [87] and what norms the actual designers may represent [6, 30].

Furthermore, moving on from this explicitly material design, by following Lessig's argument for "code" being law, we can argue that this normativity is not only expressed through the robots' embodied appearances—for example if they are gendered or not [71]—but also in how their abilities are following from programming in relation to training data. And, even more so, given the focus on adaptive technologies in this analysis, what normative aspects can be picked up in how a robot talks and positions itself in the social relation to the humans interacting with it. For example, what normative positioning do the virtual agents in the UNESCO report mentioned above perform? They are designed to be female, and to respond with a submissive tonality—"I'd blush if I could". In this case, it is not so much about physically *embodied* normative expression of gendered attributes—at least not in the same direct sense as with gendered social robots—but it is still an expression of a normative structure represented by code, and possibly, an adaptive approach enabled by the coding that may reproduce harmful stereotypes in its interaction with users. All of which stresses a need for an awareness in social robotics of what this mirroring means and may lead to.

2.1.3 Bridging Socio-Legal Studies and Gender Studies through Feminist Legal Theory

Before the next Sect. 2.2, it is worth highlighting the linkages between gender-related struggles and legal and social norms. The wave metaphor is a popular tool for telling the recent history of feminist struggles. The division brings attention to successful achievements in struggles for women's rights in predominantly Western contexts, yet the classification has been criticised by scholars for adhering to a logic of progress and fixing specific types of struggles or approaches to particular decades [88]. By stressing the links rather than the discontinuities between different theoretical frameworks, it is possible to situate the history of feminism in an understanding that brings attention to feminism's history as a series of ongoing contests and relationships, rather than discontinuities, within which feminism is characterised by heterogeneity, tension and critique [89]. While first-wave feminism was a movement recognised for its struggles to achieve legal recognition in relation to women's own personhood (e.g. not having the right to own property, or the right to vote), second-wave feminism, in turn, brought attention to the private sphere as a political arena, highlighting issues such as women's unpaid labor, in terms of domestic labor, birth control and economic empowerment [90].

While tensions and disagreements around questions of inclusion and exclusion in feminism and in society always have been a central part of struggles for gender rights, within third wave feminism, the notion of intersectionality established a conceptual framework for recognising multiple, interacting axes of power [91] and third wave feminism came to be known as the era in which the inclusivity of all women were recognised [92, 93]. The above discussion shows the importance of recognising the existence of asymmetric power relations in society, to grasp the impact of social norms on women's everyday lives and the possibilities and limitations of legal instruments for establishing justice. Overlooking these dynamics within the sphere of engineering could reproduce problematic discourses, such as issues of universalising rather than contextualizing data [94, p.8]. In the section below, we will concentrate more on the gender studies aspect within the realm of algorithms, fairness, the social, and the human.

2.2 Gender Studies

Broadly defined, gender is the social meaning ascribed to a body (ie a body identified as female or male or non-binary presenting)[30]. When distinguishing sex from gender, feminist scholars have rejected explanations of gender derived from sex [91, 95]. By illuminating how gendered identities are reproduced through social institutions, such as for example the family, education or media, feminist scholars have challenged the very idea of an essential sex [51, 94, 96, 97]. In addition, as black feminist theorists have demonstrated [29], dynamics of gender always intersect with other social categories such as processes of racialisation, sexuality and class.

2.2.1 Fairness and the Need for Situatedness

Today, efforts to achieve fairness in applied AI-systems have become popular (cf. [98]), and a debate about the limits and possibilities of such attempts has emerged among feminist, anti-racist and gender scholars of AI, who stress that inequality is reflected and amplified in algorithmic systems in ways that statistical methods only partially can address [18]. To take action against the building of devices, platforms and systems that serve to propagate sexism and racism, scholarly interventions from feminist and anti-racist traditions of knowledge highlight the need for a shift in existing theorisations of "algorithmic" fairness (which offers many definitions, but in general deals with questions of bias in relation to machine learning-based prediction, cf. [99]). With the notion 'the impossibility of fairness', scholars argue that, in an unequal society, decisions rooted in formal equality will still produce substantive inequality [100]. These ongoing debates challenge the focus on fairness as a property of

the technology itself and opens a debate regarding AI systems and existing relations of domination and oppression [101, 102]. Within this debate, scholars have highlighted the problems that can appear with attempts at moving beyond classification. They have shown that such approaches can fail to account for the harms that surface in the design of the system itself. For example, automated gender recognition systems encode the notion that gender is a binary, immutable, physiological form of identity [103]. But the reproduction of such assumptions in larger systems can give in harmful experiences among the users of technology. For example, as Sasha Constanza-Chock describes their experience in airport security:

> As a non-binary trans* femme, I present a problem not easily resolved by the algorithm of the security protocol. Sometimes, the agent will assume I prefer to be searched by a female agent; sometimes, a male. Occasionally, they ask for my preference. Unfortunately, "neither" is an honest but unacceptable response. Today, I'm particularly unlucky: a nearby male-presenting agent, observing the interaction, loudly states "I'll do it!" and strides over to me. I say, "Aren't you going to ask me what I prefer?" He pauses, then begins to move toward me again, but the female-presenting agent who is operating the scanner stops him. She asks me what I prefer. Now I'm standing in public, flanked by two TSA agents, with a line of curious travelers watching the whole interaction. Ultimately, the male-presenting agent backs off and the female-presenting agent searches me, making a face as if she's as uncomfortable as I am, and I'm cleared to continue on to my gate. [104, p.4].

Using the notion of *algorithmic oppression* [17], researchers bring to light the multiple, mundane ways in which (what we call) adaptive technologies negatively affect the lives of women, trans people, people of colour and people with disabilities (for a list of selected examples, see Myers West [105]). Showing that the effects of "algorithmic oppression" are not evenly distributed, research highlights that women and gender minorities, people of colour, people of lower socioeconomic status and people with disabilities are more strongly affected by them, and especially those whose identities lie at the intersection between several of categories [20, 25, 106–108]. Recognising these problematic implications, Sarah Myers West [52] suggests that we should move from individualised notions of "algorithmic" fairness to approach, instead, algorithmic modeling as *situated practice*.

To approach algorithmic modeling as situated practice would involve an understanding of fairness (or what in this contexts sometimes is referred to as "social justice") that starts from real-world problems of domination and oppression, rather than abstract models or categorisations [109]. Such an approach understands fairness as a property of the social context within which the problem emerges, rather than approaching it as property of technical tools [110]. Such an understanding recognises issues of decision-making, division of labor, and culture, as having an impact on fairness, despite the fact that they often are ignored in philosophical as well as technical discussions [109].

2.2.2 Resituating HRI: Who Profits?

One key aspect highlighted by feminist technoscientists is the ambition not to attribute an exact human-like agency to the robotic other, but to make the more-than-human entity intelligible within human–robotic interaction. How is it possible to allow such new patterns to emerge? Here, studies which are already classics in the field have pointed at problems with the decontextualised nature of visions or promises of robotic interaction, for example within notions of the machine-worker, and argued that the technologies cannot be developed outside of the power relations that shape the different societal spheres of production, consumption and reproduction [111–113]. Yet, reflecting a shift that recently has taken place, from focusing on rational-cognitive processes and problem-solving, to emphasising socio-emotional interaction, today, scholars focus on two main issues: sociality and emotionality when they explore the capacities for developing mutual understanding in relationships between humans and robots, including both physical robots and virtual chatbots [114, 115]. Some HRI developers have taken the relationship between infant-caregiver as a rolemodel for such exploration, tying the design to a developmental trajectory and to existing forms of relationality between humans. Nonetheless, researchers problematise the fact that such rolemodels often lack the social and cultural meaning of the figure of both the child and the role of the caregiver [116, 117]. Typically female-marked modes of bonding may downplay the symbolic ordering of the social, and risk to naturalise feminine traits as necessary for giving care [34, 115, 118].

Importantly, while these debates focus on how certain design decisions determine capacities of the robot, they do so by stipulating what social interaction means. Thereby they also define human-to-human interaction. In this context, feminist technoscience scholars [119, 120] challenge the de-contextualisation within which much of these developments have taken place, and suggests to re-situate these processes within specific arrangements of power by asking Who profits?, bringing in questions of responsibility for such "engineering of the social" [121, p.37]. As they connect the discussion of human and robot relationships to the societal division of labor and to existing divides between production, consumption and reproduction, this scholarly debate contributes with a significant re-contextualisation of present and

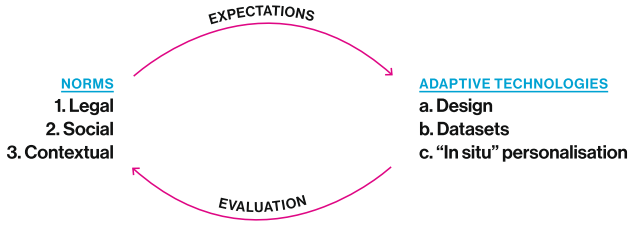

Fig. 1 Norms-in-the-loop: The Mirroring of Norms

future forms of work. Still, however, to a large extent today, the engineering of social and emotional robotic assistants are tied to notions of humanness based on certain gendered, racial, and occupational norms. One example here is the robot Nadine, assembled in 2013, created by Nadia Magnenat Thalmann (IMI Singapore) [122]. Moving from receptionist to social worker, Nadine is modelled after her creator, a White, college-educated, middle-class woman, and represents the fields of work that this robot will take over, in the realm of social work as well as the broader educational system. While such engineering of sociality involves both defining a problem in society—like the need of more workers in social work, teachers, health care workers—and delivering a possible technical solution to this, researchers have problematised the fact that the human labour forces that shape the basis for designing the robots, are coded as universal, despite the ways in which these labour forces are infused with gendered, racial and socio-economic power relations and stereotypes.

3 A Theoretical Framework

In this article, we outline a theoretical backdrop from SoL and Gender Studies in order to propose a conceptual framework on *socio-legal robotics*. Figure 1 describes a relationship between norms and adaptive technologies. Based on the presentations above, we divide norms into three categories:

1. *Legal norms*, that is, formalised normative claims. While there is a rich literature on both the relationship to technological innovation, as well as the relationship to social norms, we mainly focus the following category:
2. *Social norms*. While we acknowledge the vast literature on how to define norms, as well as the possible benefits in terms of acceptability that robots adapting to social norms may give, we focus problematic aspects of harmful social norms and stereotypes that risk being reproduced and amplified through adaptive technologies. Lastly, in lack of a more precise terminology, we include
3. *Contextual norms*, which means an emphasis on the need for contextuality and situatedness of how to study and understand adaptive technologies as played out and dependent on norms in various contexts.

When it comes to the normative mirroring, the *norms-in-the-loop*, so to speak, we roughly divide the adaptive technologies into

a. *Design*, pointing to both how robots can be embodied and e.g. gendered, and may reveal biases by the designers, and reproduce norms in a very material sense. Secondly, we point to
b. *Datasets*, which is acutely shown in recent debates on fairness in AI, where the sampling may either be biased in that it does not represent actual distributions in society, or biased in the sense that is represents an unfair society as such, which raises normative questions for those that automate and reproduce this unfairness in data-dependent design, algorithmic decision-making or AI-modelling. Lastly, we point to
c. *"In situ" personalisation*, in order to acknowledge normative questions for personalised robotics or other adaptive technologies that may learn from single individuals and individualise their feedback—perhaps well exemplified by a personalised chatbot used by an individual with self-harming behaviour. What is a fair position for the chatbot? Regardless, the position will be normative.

The interplay between norms and adaptive technologies are inspired by Iyad Rahwan's account on society-in-the-loop [28], clearly pointing to the societal expectations that evaluate and shape new technologies and methods like machine learning and AI-systems. For robots, this has been referred to as a sort of mutual shaping [123, 124], here framed as mirroring of norms.

3.1 Discussion

3.1.1 Norms

Firstly, and as outlined in Sect. 2.1 above, the interplay and friction between legal and social norms (see Fig. 1) has been studied in detail over a long time (cf. [61, 68, 69]). One the one hand, there are much regulation that firmly depends on supportive social informal norms [65] (as a social basis for law [61]). On the other, there are social structures and community norms that themselves may be expressing violent or harmful unfairness, such as discrimination, that law is set to try to come to terms with through regulation [23]. In Fig. 1 we also include contextuality, as a way to acknowledge the need pointed to in Sects. 2.2.1 and 2.2.2 above to regard context and situatedness when analysing social challenges in relation to adaptive technologies. This overlaps social norms to some extent, since many social expectations are triggered by or linked to certain contexts (cf. [125]), but is also a way to acknowledge domain specific norms, such as medical ethics or professional conduct (cf. [7]). For HRI developers to bring

in social robots in certain settings, we argue, developers must understand the context in which they bring the robot [126, p.150], and what social implications this may have [45].

3.1.2 Adaptive Technologies

Secondly, bearing HRI in mind, we distinguish between three levels of adaptive technologies found to the right in Fig. 1. Of these, the overarching, on *design*, is of course very general, but a way to point to material aspects of robotic design, that we have pointed to as being normative in the sense of "code-as-law" in Sect. 2.1.2. Designs that gender robots are, for example, a clear example of the mirroring of norms, which at the same time may have many implications [10, 127]. The reason to include *datasets* as an adaptive technology in Fig. 1 is leaning on the recent and ongoing debates on bias and fairness in machine learning and AI, such as pointed to in the gender shades study [20]. However, there is an interesting duality in how to look at biased datasets pointed to by Larsson [26]. On the one hand, they may be biased in how they represent society, i.e. lacking training data for certain groups, cultures, regions or other phenomena—which may lead to a sort of algorithmic oppression [17]. The solution proposed to remedy this is often to collect more data, attempting to make the datasets more representative of society. On the other hand, which stresses the link to social norms further, one can also picture cases where the data may represent society fairly well, but society as such is skewed and unfair in how it distributes power, access to privileges such as work, education etc. In these cases, the data may actually be mirroring social structures or communities as such, but the challenge for adaptive technologies is found in that they reproduce or amplify violent, discriminatory, sexist or racist sentiment. For example, the datasets utilised for training ML-based prediction in image recognition or other classification algorithms can be heavily reliant on how and by whom the annotation is done. This has been problematised in terms of the construction of race and gender in an analysis of a large annotated database called ImageNet [80]. With regards to training data, this creates another type of challenge than non-representative datasets for AI-systems and robotic design, which is normative in relation to how to intervene or more actively scrutinise and engage in what it is a particular adaptive technology is supposed to do. At worst, and without awareness of how to deal with this set of problems, robots may merely reproduce also the types of social structures that would be harmful and unfair.

With the last category in Fig. 1, regarding the *in situ personalisation*, we point to individualised personalisation in the interaction of humans and robots (cf. [128]). By drawing from personalisation in other automated services like social media or artificial agents like the Replika chatbot [129], we see yet another type of interaction and adaptation to address with regards to adaptive technologies and norms (cf. [37, 38, 130, 131]).

3.1.3 On the Mirroring of Norms

By pointing to a theoretical framework that includes the notion of norms as measurable facts, we hope to provide with awareness of the role of informal social structures also in robotic design and HRI. To be clear, the existence of normative facts, that is, social norms, by no means mean that they are inherently fair and desirable for robotic "alignment". It only means that there is an existing structure linked to social expectations, that may explain certain behaviour. In fact, these social structures may be linked to behaviour that can be useful for understanding social interplay—and in some cases even increase acceptability of robots in some groups—but at the same time at worst also be harmful, divisive, misogynist and sexist. Social norms are not necessarily fair, so to speak.

The dynamics of "personalised" social robotics, set to adapt to an individual user, should therefore arguably not only be measured in terms of acceptability, as is common in HRI, but also stress a critical scrutiny of what norms are reproduced or amplified in this adaptive relationship. For example, robot design or learning techniques that aim to mimic human behavior are argued to not necessarily guarantee *fair* behavior [59]. Here, one may ask what it means to include aspects of social structures, for example what informal normative structures and human expressions related to gender, ethnicity, age, culture, language, as part in robotic learning. How should we detect and understand unfairness within this frame? To be able to contribute to this type of knowledge and its range of technological, methodological and theoretical dimensions, a research programme not only needs to include competence on the traditional strands in social robotics combined with aspects of the research fronts of computational AI-research, but also the theoretical underpinnings of disciplines that since-long have studied such social structures and their implications—for example SoL and Gender Studies.

We do not offer an answer to how to best handle or "solve" robotic mirroring of social norms, but stress a need for more awareness of this phenomenon. Informed by a socio-legal research paradigm, what values that are to be regarded as the achievable ones, in general and for various social contexts, is a core challenge of all communities and societies. In this context, feminist technoscientist scholars pointed to above intervenes into the techno-deterministic approaches which currently locates the robotic imaginary in an either-or position between a "utopian, welcoming position or a dystopian, resistant position", as they seek new approaches that could open up for the development of understandings of "socially-just kinds of human–robot co-habitation" [121, 132]. Further,

within the fair-ML community, a key goal is to develop ML and automated systems that can achieve fairness in social and legal settings. However, scholars have shown that the concepts used to define notions of fairness and discrimination renders technical interventions "ineffective, inaccurate and sometimes dangerously misguided when they enter the societal context that surrounds decision-making systems" [110, p.59], mainly because such concepts fail to consider how the social context intermesh with technology in different forms. A change of focus of designs, some scholars argue, would mitigate the traps, for instance by refocusing AI designs in terms of processes instead of solutions, and by including social actors and different stakeholders into the abstraction boundaries, rather than being limited to purely technical dimensions [110].

Lastly, many of the examples given above relate to non-embodied examples of adaptive technologies—such as face-recognition, language-models or virtual agents. These do however arguably show what functionalities that are likely to become included in robotic, embodied, applications. This means that the social robotics field can learn from examples, mistakes and problematic cases from the non-embodied but adaptive AI-systems. It also means, following the material and coded architectures, that the social norms are not only to be learned in data-collection, but can also, obviously, be expressed in the materialities of design as such.

3.1.4 Recontextualising the Decontextualised

Social Robotics and HRI as a field is young [34, p.3], and research is continuously finding ways to bring social robots in society, as seen with the yearly HRI Conferences. Following the suggestions of feminist and anti-racist scholars, algorithmic modeling as a situated practice can provide a more robust way to hold the institutions creating and deploying AI-systems to account by affirming, rather than downplaying, difference [105, 110, 133, 134]. As Sarah Myers West highlights [105], some examples of such ambitions already exist, for instance in the Feminist Data Manifest-No [133], as well as in the calls of scholars upon technical designers to redraw their abstraction boundaries to include social actors [110]. Further, researchers push for ways in which a decolonial critical approach can be embedded in technical practice [134] to overcome the structural barriers that inhibit the development of a feminist "AI from below" [105]. For example, a systematic review of sampling in HRI research found it to be lacking diversity [87]. Another study explored the impact of overlooking gender and sex consideration in robot design on users [135]. This means that the "who" and "for whom" of HRI research are key factors to acknowledge since a lack of awareness risks propagating universalist claims for phenomena that are not universal. This is well in line with how we argue for a need for *recontextualising the decontextualised visions of technologies*, often presented as universal, despite the ways they may be gendered and infused with racial and socio-economic power relations and stereotypes.

Lastly, looking to the future, as social robotics becomes commonplace, it will not only be highly entangled with social norms and the complexities of interaction with humans, but also embedded in commercial strategies, datafied and shaped to fit business models of various sorts. Earlier shifts in technology-development can teach some of the implications of that transition. The early Internet that Lessig and others saw as a distributed and layered enabler of innovation has now morphed into a more *regionalised* and *platformised* sociotechnical construct, highly dependent on commodification of data, feeding the underlying business models (cf. [83, 136]). Often in the shape of ad-tech, or geopolitical struggles of dominion and control. As social robots increasingly become commonplace—and data-collecting, internet-connected entities—this field too will likely meet all sorts of similarly entangled issues relating to power, markets, business models and governmental control. This further calls for what feminist technoscience scholars [119, 120] suggest in terms of re-situating these processes within specific arrangements of power by asking: who profits? Any technodeterministic approach likely needs to be situated and challenged, stripped from universalistic attributes and scrutinised for what it actually is reproducing, for what reason and for whom.

4 Conclusions

In this article, we outline a theoretical backdrop from SoL and Gender Studies in order to propose a conceptual framework on *socio-legal robotics*. Here we seek to combine these disciplinary insights with HRI in an attempt to accommodate for what we see as emerging concerns in social robotics. That is, to be able to theorise and understand how to deal with the fact that the underlying technologies increasingly are becoming adaptive of the social interplay that includes social norms and stereotypes, here with a particular focus on gender. On the one hand, there are studies pointing to the usefulness of mirroring human traits in robots when striving for acceptability in human users [7, 46–48]. On the other, there are risks of mirroring social norms relating to for example gender—including stereotyping, sexism, and racism—pointed to in critical AI-research [17, 20, 25, 52, 75]. In short, we have:

- Proposed a theoretical basis of *socio-legal robotics*, primarily drawn from the realms of social sciences that underpins both Sociology of Law and Gender Studies. This focuses on social norms, relates to legal norms, while at the same time emphasises the need for the inclusion of context and situatedness.

- Proposed a framework that distinguishes between (i) design, (ii) datasets, and (iii) in situ personalisation as three distinct aspects of *adaptive technologies*. They adapt and mirror norms in different ways, which includes perceptions of the persons designing, issues of biased datasets as well as normative challenges inherent in technologies that adapt and personalise on an individual level.
- Related the theoretical framework to already established notions found in social robotics as a field, such as acceptability and personalisation; with particular focus on the adaptive interplay between AI-supported technologies and human social structures. If a robot adapts to and "learns" social norms, we point not only to potentially beneficial aspects of acceptability for certain users but problematise in terms of risks for reproducing or amplifying harmful, sexist, racist or otherwise deeply problematic stereotypes.
- Developed an account on fairness as a contextualised and situated practice in human–robot interaction, in order to be able to detect and avoid undesired or unfair aspects of robotic mirroring of social norms.

Lastly, the main argument depends on a theoretical understanding of societal unfairness. This opens for contributions from critical social sciences, to the already interdisciplinary domains of HRI, as advancements in adaptive technologies are incorporated into social robotics.

Acknowledgements A special thank you to Katie Winkle for drawing our attention to the in situ concept.

Author Contributions All authors contributed to the study conception and design. This is a highly conceptual and argumentative paper, based on three research disciplines, of which the four authors relate to differently. The socio-legal foundation is primarily performed by SL and LT, the Gender Studies foundation primarily by ML and the social robotics foundation is primarily performed by GC. The first draft of the manuscript was written by SL and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Open access funding provided by Lund University. This work was supported by (i) the Wallenberg AI, Autonomous Systems and Software Program-Humanities and Society (WASP-HS) funded by the Marianne and Marcus Wallenberg Foundation and the Marcus and Amalia Wallenberg Foundation, and (ii) Swedish Research Council, grant no 2018-03869.

Declarations

Conflict of interest: The authors declare that they have no conflict of interest.

Ethics approval: Not applicable.

References

1. Breazeal C (2003) Toward sociable robots. Robot Auton Syst 42(3–4):167–175
2. Bartneck C, Forlizzi J (2004) A design-centred framework for social human-robot interaction. Robot and human interactive communication. IEEE Xplore, Roman, pp 591–594
3. Dautenhahn K, Billard A (1999) Bringing up robots or-the psychology of socially intelligent robots: from theory to implementation. In: Third Annual Conference on Autonomous Agents, pp. 366–367
4. Fong T, Nourbakhsh I, Dautenhahn K (2003) A survey of socially interactive robots. Robot Auton Syst 42(3–4):143–166. https://doi.org/10.1016/S0921-8890(02)00372-X
5. Fosch-Villaronga E, Lutz C, Tamò-Larrieux A (2020) Gathering expert opinions for social robots' ethical, legal, and societal concerns: findings from four international workshops. Int J Soc Robot 12(2):441–458. https://doi.org/10.1007/s12369-019-00605-z
6. Tanqueray L, Larsson S (2023) What norms are social robots reflecting? a socio-legal exploration on hri developers. In: Social Robots in Social Institutions, pp. 305–314. Proceedings of the 2022 Robophilosophy Conference in Helsinki
7. Tanqueray L, Paulsson T, Zhong M, Larsson S, Castellano G (2022) Gender fairness in social robotics: exploring a future care of peripartum depression. In: ACM/IEEE International Conference on Human-Robot Interaction (HRI 2022). Association for Computing Machinery (ACM)
8. Šabanović S (2014) Inventing Japan's 'robotics culture': the repeated assembly of science, technology, and culture in social robotics. Soc Stud Sci 44(3):342–367
9. Robertson J (2010) Gendering humanoid robots: robo-sexism in Japan. Body Soc 16(2):1–36
10. Winkle K, Jackson RB, Melsión GI, Brščić D, Leite I, Williams T (2022) Norm-breaking responses to sexist abuse: a cross-cultural human robot interaction study. In: Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction, pp. 120–129
11. Ramis S, Buades JM, Perales FJ (2020) Using a social robot to evaluate facial expressions in the wild. Sensors 20(23):6716
12. James J, Balamurali B, Watson CI, MacDonald B (2021) Empathetic speech synthesis and testing for healthcare robots. Int J Soc Robot 13:2119–2137
13. Panesar S, Cagle Y, Chander D, Morey J, Fernandez-Miranda J, Kliot M (2019) Artificial intelligence and the future of surgical robotics. Ann Surg 270(2):223–226
14. Strengers Y, Kennedy J (2021) The smart wife: Why Siri, Alexa, and other smart home devices need a feminist reboot. MIT Press, Cambridge
15. Winkle K, McMillan D, Arnelid M, Balaam M, Harrison K, Johnson E, Leite I (2023) Feminist human-robot interaction: disentangling power, principles and practice for better, more ethical hri. In: Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI)

16. O'Neil C (2016) Weapons of math destruction: how big data increases inequality and threatens democracy. Broadway Books, Panaji
17. Noble SU (2018) Algorithms of oppression. New York University Press, New York
18. Benjamin R (2019) Race after technology: abolitionist tools for the new Jim code, p. 172. Polity
19. Costanza-Chock S (2020) Design justice: community-led practices to build the worlds we need. The MIT Press, Cambridge
20. Buolamwini J, Gebru T (2018) Gender shades: intersectional accuracy disparities in commercial gender classification. In: Conference on Fairness, Accountability and Transparency, pp. 77–91. PMLR
21. Datta A, Tschantz MC, Datta A (2015) Automated experiments on ad privacy settings. Proc Priv Enhanc Technol 1:92–112
22. Neff G, Nagy P (2016) Talking to bots: symbiotic agency and the case of Tay. Int J Commun 10:17
23. Larsson S (2019) The socio-legal relevance of artificial intelligence. Droit et société 103(3):573–593
24. Susskind J (2018) Future politics: living together in a world transformed by tech. Oxford University Press, Oxford
25. Eubanks V (2018) Automating inequality: how high-tech tools profile, police, and punish the poor. St. Martin's Press, New York
26. Larsson S (2021) AI in the EU: ethical guidelines as a governance tool. The European Union and the Technology Shift, 85–111
27. Mandel GN (2020) Regulating emerging technologies, 361–378
28. Rahwan I (2018) Society-in-the-loop: programming the algorithmic social contract. Ethics Inf Technol 20(1):5–14
29. Crenshaw K (1991) Mapping the margins: intersectionality, identity politics, and violence against women of color. Stanford Law Rev 43(6):1241–1299
30. Criado Perez C (2020) Invisible women: data bias in a world designed for men. Penguin Random House, New York
31. Wajcman J (2004) TechnoFeminsm. Polity Press, Oxford
32. Chernova S, Thomaz AL (2014) Robot learning from human teachers. Synth Lect Artif Intell Mach Learn 8(3):1–121
33. Gao Y, Sibirtseva E, Castellano G, Kragic D (2019) Fast adaptation with meta-reinforcement learning for trust modelling in human-robot interaction. In: Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2019)
34. Hakli R, Seibt J (2017) Sociality and normativity for robots. Springer, Berlin
35. Akalin N, Loutfi A (2021) Reinforcement learning approaches in social robotics. Sensors 21(4):1292
36. Jones A, Castellano G (2018) Adaptive robotic tutors that support self-regulated learning: a longer-term investigation with primary school children. Int J Soc Robot 10(3):357–370
37. Churamani N, Anton P, Brügger M, Fließwasser E, Hummel T, Mayer J, Mustafa W, Ng HG, Nguyen TLC, Nguyen Q, et al (2017) The impact of personalisation on human-robot interaction in learning scenarios. In: Proceedings of the 5th International Conference on Human Agent Interaction, pp. 171–180
38. Rossi A, Rossi S (2021) Engaged by a bartender robot: recommendation and personalisation in human-robot interaction. In: Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization, pp. 115–119
39. Lucy L, Bamman D (2021) Gender and representation bias in gpt-3 generated stories. In: Proceedings of the Third Workshop on Narrative Understanding, pp. 48–55
40. Bousmalis K, Vezzani G, Rao D, Devin C, Lee AX, Bauza M, Davchev T, Zhou Y, Gupta A, Raju A, et al (2023) Robocat: a self-improving foundation agent for robotic manipulation. arXiv preprint arXiv:2306.11706
41. Ayub A, Wagner AR (2020) What am i allowed to do here?: Online learning of context-specific norms by pepper. In: International Conference on Social Robotics, pp. 220–231. Springer
42. Paiva A, Leite I, Ribeiro T (2014) 21 emotion modeling for social robots. The Oxford handbook of affective computing, 296
43. Tomic S, Pecora F, Saffiotti A (2018) Norms, institutions, and robots. arXiv preprint arXiv:1807.11456
44. Winkle K, Melsión G.I, McMillan D, Leite I (2021) Boosting robot credibility and challenging gender norms in responding to abusive behaviour: a case for feminist robots. In: Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction. HRI '21 Companion, pp. 29–37. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3434074.3446910
45. Serholt S, Ljungblad S, Ní Bhroin N (2021) Introduction: special issue-critical robotics research. Springer, berlin
46. Riek L, Rabinowitch T, Chakrabarti B, Robinson P (2009) Empathizing with robots: Fellow feeling along the anthropomorphic spectrum. In: Proceedings of the 2009 International Conference on Affective Computing and Intelligent Interaction
47. Andrist S, Mutlu B, Tapus A (2015) Look like me: matching robot personality via gaze to increase motivation. In: ACM Conference on Human Factors in Computing Systems
48. Paetzel M, Perugia G, Castellano G (2020) Persistence of first impressions: The effect of repeated interactions on the perception of a social robot. In: 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI2020)
49. Mori M, MacDorman KF, Kageki N (2012) The uncanny valley. Robot Autom Mag 19(2):98
50. Thepsoonthorn C, Ogawa K-i, Miyake Y (2021) The exploration of the uncanny valley from the viewpoint of the robot's nonverbal behaviour. Int J Soc Robot 13(6):1443–1455
51. D'Ignazio C, Klein LF (2020) Data feminism. The MIT Press, Cambridge, p 328
52. West M, Kraut R, Ei Chew H. I'd blush if i could: closing gender divides in digital skills through education
53. Eyssel F, Hegel F (2012) (s)he's got the look: gender-stereotyping of social robots. J Appl Soc Psychol 42(9):2213–2230
54. Bryant D, Borenstein J, Howard A (2020) Why should we gender? the effect of robot gendering and occupational stereotypes on human trust and perceived competency. In: Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction. HRI '20, pp. 13–21. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3319502.3374778
55. Perugia G, Guidi S, Bicchi M, Parlangeli O (2022) The shape of our bias: Perceived age and gender in the humanoid robots of the abot database. In: Proceedings of the 2022 17th ACM/IEEE International Conference on Human-Robot Interaction
56. Jackson R.B, Williams T, Smith N (2020) Exploring the role of gender in perceptions of robotic noncompliance. In: Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction. HRI '20, pp. 559–567. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3319502.3374831
57. Winkle K, Melsion G, Leite I, Balaam M, McMillan D, Arnelid M, Harrison K, Johnson E (2021) Feminism x social robotics. GenR Workshop. GENDERING ROBOTS: Ongoing (Re)configurations of Gender in Robotics
58. Hurtado J.V, Mejia V (2022) Feminist perspective on robot learning processes. arXiv preprint arXiv:2201.10853
59. Hurtado JV, Londoño L, Valada A (2021) From learning to relearning: a framework for diminishing bias in social robot navigation. Front Robot AI 8:650325
60. IEEE (2019) Ethically aligned design. A vision for prioritizing human well-being with autonomous and intelligent systems

61. Cotterrell R (1992) The sociology of law: an introduction. Oxford University Press, Oxford
62. Durkheim E (1982) Rules of sociological method. Simon and Schuster, New York
63. Banakar R (2015) Driving culture in Iran: law and society on the roads of the Islamic republic. Bloomsbury Publishing, London
64. Hydén H, Svensson M (2008) The concept of norms in sociology of law. Contrib Soc Law Remarks Swed Horiz 53:129–146
65. Hydén H (2022) Sociology of law as the science of norms. Taylor & Francis, Milton Park
66. Svensson M, Larsson S (2012) Intellectual property law compliance in Europe: Illegal file sharing and the role of social norms. New Media Soc 14(7):1147–1163
67. Ellickson RC (1991) Order without law. Harvard University Press, Harvard
68. Ehrlich E (1913) Grundlegung der Soziologie des Rechts. Duncker & Humblot, Munich
69. Pound R (1910) Law in books and law in action. Am L Rev 44:12
70. Bicchieri C, Muldoon R, Sontuoso A (2018) Social Norms. In: Zalta EN (ed) The Stanford Encyclopedia of Philosophy, Winter 2018 edn Metaphysics Research Lab. Stanford University, Stanford
71. Wessel M, Ellerich-Groppe N, Schweda M (2021) Gender stereotyping of robotic systems in eldercare: an exploratory analysis of ethical problems and possible solutions. Int J Soc Robot. https://doi.org/10.1007/s12369-021-00854-x
72. Birhane A, Prabhu VU, Kahembwe E (2021) Multimodal datasets: misogyny, pornography, and malignant stereotypes. arXiv preprint arXiv:2110.01963
73. Stewart R, Wright B, Smith L, Roberts S, Russell N (2021) Gendered stereotypes and norms: a systematic review of interventions designed to shift attitudes and behaviour. Heliyon 7(4):06660
74. Wang C, Wang K, Bian A, Islam R, Keya KN, Foulds J, Pan S (2021) User acceptance of gender stereotypes in automated career recommendations. arXiv preprint arXiv:2106.07112
75. Dastin J (2018) Amazon scraps secret AI recruiting tool that showed bias against women. Ethics of data and analytics. Auerbach Publications, Boca Raton, pp 296–299
76. Gabriel I (2020) Artificial intelligence, values, and alignment. Mind Mach 30(3):411–437
77. Whittaker M, Alper M, Bennett C.L, Hendren S, Kaziunas L, Mills M, Morris M.R, Rankin J, Rogers E, Salas M, et al (2019) Disability, bias, and AI. AI Now Institute. 8
78. Hydén H (2020) Sociology of digital law and artificial intelligence. In: Pibá J (ed) Research handbook on the sociology of law. Elgar, Cheltenham
79. Raghavan M, Barocas S, Kleinberg J, Levy K (2020) Mitigating bias in algorithmic hiring: evaluating claims and practices. In: Proceedings of the 2020 Conference on fairness, accountability, and transparency, pp. 469–481
80. Crawford K (2021) The atlas of AI. Yale University Press, London
81. Lessig L (2003) Law regulating code regulating law. Loy U Chi LJ 35:1
82. Lessig L (2006) Code: and other laws of cyberspace. Basic Books, New York
83. Larsson S (2021) Putting trust into antitrust? competition policy and data-driven platforms. Eur J Commun 36(4):391–403
84. Käll J (2022) Posthuman property and law: commodification and control through information, smart spaces and artificial intelligence. Taylor & Francis, Milton Park
85. Kozuka S (2019) A governance framework for the development and use of artificial intelligence: lessons from the comparison of japanese and european initiatives. Unif Law Rev 24(2):315–329
86. Leenes R, Lucivero F (2014) Laws on robots, laws by robots, laws in robots: regulating robot behaviour by design. Law Innov Technol 6(2):193–220
87. Seaborn K, Barbareschi G, Chandra S (2023) Not only weird but uncanny? a systematic review of diversity in human-robot interaction research. Int J Soc Robot. https://doi.org/10.1007/s12369-023-00968-4
88. Hemmings C (2005) Telling feminist stories. Fem Theory 6(2):115–139
89. Tong R (2009) Feminist thought: a more comprehensive introduction. Westview Press, Nashville
90. Samuels H (2013) Feminist legal theory. Hart, London
91. De Beauvoir S (1949) The second sex. Rowman & Littlefield Publishers, Lanham
92. Nielsen R, Tvarnø CD (2012) Scandinavian women's law in the 21st century. Djøf Forlag, København
93. Draude C, Hornung G, Klumbytė G (2022) Mapping data justice as a multidimensional concept through feminist and legal perspectives. In: New Perspectives in Critical Data Studies: The Ambivalences of Data Power, pp. 187–216. Springer, Berlin
94. Ahmed S (2004) The cultural politics of emotion. Edinburgh University Press ltd, Great Britain
95. Rubin G (1975) The traffic in women: notes on the political economy of sex
96. Butler J (1988) Performative acts and gender constitution: an essay in phenomenology and feminist theory. Theatr J 40(4):519–531
97. Collins PH (1998) Fighting words: black women and the search for justice. Minnesota Press, Minneapolis
98. Wachter S, Mittelstadt B, Russell C (2021) Why fairness cannot be automated: bridging the gap between EU non-discrimination law and AI. Comput Law Secur Rev 41:105567
99. Mitchell S, Potash E, Barocas S, D'Amour A, Lum K (2021) Algorithmic fairness: choices, assumptions, and definitions. Ann Rev Stat Appl 8:141–163
100. Green B (2020) The false promise of risk assessments: epistemic reform and the limits of fairness. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 594–606
101. Hutchinson B, Mitchell M (2019) 50 years of test (un) fairness: lessons for machine learning. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 49–58
102. Barocas S, Selbst AD (2016) Big data's disparate impact. Calif L Rev 104:671
103. Keyes O (2018) The misgendering machines: Trans/HCI implications of automatic gender recognition. Proceedings of the ACM on human-computer interaction 2(CSCW), 1–22
104. Constanza-Chock S (2020) Design justice. MIT Press, Cambridge
105. West SM (2020) Redistribution and rekognition: a feminist critique of algorithmic fairness. Catal Fem Theory Technosci. 6(2)
106. Raji ID, Buolamwini J (2019) Actionable auditing: investigating the impact of publicly naming biased performance results of commercial AI products. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pp. 429–435
107. Ali M, Sapiezynski P, Bogen M, Korolova A, Mislove A, Rieke A (2019) Discrimination through optimization: How facebook's ad delivery can lead to biased outcomes. Proceedings of the ACM on Human-Computer Interaction 3(CSCW), 1–30
108. Vincent J (2016) Twitter taught microsoft's ai chatbot to be a racist asshole in less than a day. The Verge. 24
109. Young IM (1990) Justice and the politics of difference. Princeton University Press, Princeton
110. Selbst AD, Boyd D, Friedler SA, Venkatasubramanian S, Vertesi J (2019) Fairness and abstraction in sociotechnical systems. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 59–68
111. Heßler M (2001) "Mrs. Modern Woman": zur Sozial-und Kulturgeschichte der Haushaltstechnisierung. Campus Verlag, Frankfurt

112. Cowan R (1983) More work for mother. Pantheon, New York
113. Cockburn C, Ormrod S (1993) Gender and technology in the making. SAGE Publications Ltd, Thousand Oaks
114. Suchman L, Suchman LA (2007) Human-machine reconfigurations: plans and situated actions. Cambridge University Press, Cambridge
115. Weber J (2005) Helpless machines and true loving care givers: a feminist critique of recent trends in human-robot interaction. J Inf Commun Ethics Soc 3:209
116. Treusch P (2015) Robotic companionship: the making of anthropomatic kitchen robots in queer feminist technoscience perspective. PhD thesis, Linköping University Electronic Press
117. Treusch P (2017) The art of failure in robotics: queering the (un) making of success and failure in the companion robot laboratory. Catalyst Femin Theory Technosci 3(2):1–27
118. Fox Keller E (2007) A clash of two cultures. Nature 445(7128):603–603
119. Star SL (1995) Ecologies of knowledge: work and politics in science and technology. Suny Press, New York
120. Haraway D (1996) Modest witness: feminist diffractions in science studies. In: Galison PL, Stump DJ (eds) The disunity of science: boundaries, contexts, and power. Stanford University Press, Stanford
121. Treusch P (2020) Robotic knitting: re-crafting human-robot collaboration through careful coboting. Transcript Verlag, Bielefeld
122. Nanyang Technological University (2012) Nadia Magnenat Thalmann (IMI Singapore). https://www3.ntu.edu.sg/imi/3d-idm/nadia_thalmann.html Accessed 2022-03-14
123. Šabanović S (2010) Robots in society, society in robots. Int J Soc Robot 2(4):439–450
124. Winkle K, Caleb-Solly P, Turton A, Bremner P (2020) Mutual shaping in the design of socially assistive robots: a case study on social robots for therapy. Int J Soc Robot 12(4):847–866
125. Rakoczy H, Schmidt MF (2013) The early ontogeny of social norms. Child Dev Perspect 7(1):17–21
126. Bhaumik A (2018) From AI to robotics: mobile, social, and sentient robots. CRC Press, Boca Raton
127. de Graaf M, Perugia G, Fosch-Villaronga E, Lim A, Broz F, Short ES, Neerincx M (2022) Inclusive hri: Equity and diversity in design, application, methods, and community. In: Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction, pp. 1247–1249
128. Senft E, Lemaignan S, Baxter PE, Bartlett M, Belpaeme T (2019) Teaching robots social autonomy from in situ human guidance. Sci Robot 4(35):1186
129. Ta V, Griffith C, Boatfield C, Wang X, Civitello M, Bader H, DeCero E, Loggarakis A (2020) User experiences of social support from companion chatbots in everyday contexts: thematic analysis. J Med Int Res 22(3):e16235
130. Reig S, Luria M, Wang JZ, Oltman D, Carter EJ, Steinfeld A, Forlizzi J, Zimmerman J (2020) Not some random agent: multi-person interaction with a personalizing service robot. In: Proceedings of the 2020 ACM/IEEE International Conference on Human-robot Interaction, pp. 289–297
131. Reig S, Luria M, Forberger E, Won I, Steinfeld A, Forlizzi J, Zimmerman J (2021) Social robots in service contexts: exploring the rewards and risks of personalization and re-embodiment. Des Interact Syst Conf 2021:1390–1402
132. Sollfrank C (2018) Die schönen kriegerinnen-technofeministische praxis im 21. Transversal, Jahrhundert Vienna
133. Cifor M, Garcia P, Cowan T, Rault J, Sutherland T, Chan A, Rode J, Hoffmann A.L, Salehi N, Nakamura L (2019) Feminist data manifest-no. Cit. on, 119
134. Mohamed S, Png M-T, Isaac W (2020) Decolonial AI: decolonial theory as sociotechnical foresight in artificial intelligence. Philos Technol 33(4):659–684
135. Fosch-Villaronga E, Drukarch H (2023) Accounting for diversity in robot design, testbeds, and safety standardization. Int J Soc Robot. https://doi.org/10.1007/s12369-023-00974-6
136. Van Dijck J, Poell T, De Waal M (2018) The platform society: public values in a connective world. Oxford University Press, Oxford

Stefan Larsson is a senior lecturer and Associate Professor in Technology and Social Change at Lund University, Sweden, Department of Technology and Society. He is a lawyer and socio-legal researcher that holds a PhD in Sociology of Law as well as a PhD in Spatial Planning. He leads a multidisciplinary research group on AI and Society that studies the impact of AI-supported technologies in various domains, such as on consumer markets, in the public sector, for health, and social robotics.

Mia Liinason is Professor of Gender Studies at Lund University. Her research is situated at the crossroad between transnational feminism, queer studies and digital technologies. She is the director of the TechnAct research cluster, exploring emerging digital cultures in global and local struggles for rights. She is co-researcher in a collaborative research program which aims to create computational tools to capture how languages, societies and cultures have changed over time. She also leads a research project exploring religious and social barriers encountered by LGBTQ+ people across diverse traditions of faith.

Laetitia Tanqueray is a PhD Candidate at the Department of Technology and Society, at Lund University Sweden. Laetitia holds law degrees (LLB and Master 1) and a Master's (MSc) in Sociology of Law. She investigates human-robot interactions (HRI) through a socio-legal lens. Her published work has mostly focused on informing HRI design, including in collaboration with HRI.

Ginevra Castellano is a Professor in Intelligent Interactive Systems at the Department of Information Technology of Uppsala University, Sweden, where she leads the Uppsala Social Robotics Lab. Her research is in the area of social robotics and human-robot interaction, addressing questions on how we can build human-robot interactions that are trustworthy, including human-robot relationship formation, robot ethics, robot autonomy and human oversight, gender fairness, robot transparency and trust, both from the perspective of developing computational skills for robotic systems, and their evaluation with human users to study acceptance and social consequences.